Estimation in truncated parameter spaces

Moors, Johannes Josephus Antonius

Publication date:
1985

estimation in truncated parameter spaces

jja moors

Estimation in Truncated Parameter Spaces

Estimation in Truncated Parameter Spaces

Proefschrift ter verkrijging van de graad van doctor in de economische wetenschappen aan de Katholieke Hogeschool Tilburg, op gezag van de rector magnificus, prof. dr. R.A. de Moor, in het openbaar te verdedigen ten overstaan van een door het college van decanen aangewezen commissie in de aula van de Hogeschool op vrijdag 21 juni 1985 te 15.00 uur precies

door

Johannes Josephus Antonius Moors

geboren te Helmond

Promotor: prof. dr. B.B. van der Genugten

Druk: Dissertatie Drukkerij Wibro, Helmond.
Omslagontwerp: Tamar Moors

Contents

Preface

The history of this thesis goes back to June 1976, when I first discussed randomized response methods as a possible research subject with Ben van der Genugten. At that time, randomized response was a fairly new interviewing technique, devised to protect the respondents' privacy - better than traditional methods could do. The estimation problem in the randomized response situtation is essentially the estimation of a binomial parameter, known in advance to lie in a given interval that is smaller than the unit interval. Three papers on this type of estimation problems appeared in the Department's internal series 'Ter Discussie'; they are the backbone of this thesis, which is hoped to have flesh and blood in addition.

The list of persons who in any way contributed to the book is impressive. Only a few of them will be mentioned explicitly. Of course, Ben van der Genugten is on top. In particular, I admire his dashing way of tackling both fundamental mathematical questions and numerical problems. For my programming problems I am much indebted to Toon van den Aker, who was willing to help at any moment of the day, and to Ad Plaisier in an earlier stage of the work. Annemiek Dikmans gave her usual show of excellent typing, while Jan Pijnenburg took very good care of the drawings. Hildegard Penn talked me out of the head that 'a kink in the cable' is a correct English expression. Of course, all remaining errors of any kind arw mine.

Finally, I have to thank my three cheerleaders: Heleen who, up to the very end, heroically hided her impatience, Sinbad who was always prepared to distract me by means of some game, and Tamar who cheered my room with her beautiful works of art.

I thank you all.

Hans Moors

Tilburg, April 23,1985

INTRODUCTION AND SUMMARY

In the empirical social sciences response bias in surveys is a great nuisance. Especially in surveys about personal, sensitive matters both refusals to answer and untruthful replies occur rather frequently. As a consequence, the problem of estimating population parameters becomes much more complicated, while the estimates obtained lack precision.

To meet this problem, WARNER 1965 developed a method of interviewing aiming at the protection of the respondent's privacy and, consequently, at the reduction - or even elimination - of response bias. The basic principle is that the respondent draws at random one out of several statements. The interviewer is not allowed to know what statement was drawn: he learns the respondent's answer 'correct' or 'false' without knowing to which statement it applies. If the two statements were chosen sensibly, neither of the two possible answers incriminates the respondent; hence more, and more truthful, cooperation may be expected.

Warner's original method concerns the incidence of some sensitive property A and makes use of the two statements:

'I possess property A'
'I do not possess property A'

One of these statements is selected by some chance mechanism and presented to the respondent. The probability that the first statement is drawn is known and equals, say, P, so that the probability of the latter is 1-P. Since for $P = \frac{1}{2}$ no inference about the incidence of A is possible and because of symmetry, $\frac{1}{2} < P \leq 1$ may be assumed. In case $P = 1$, in fact a direct question is asked. However, for $P < 1$ neither of the two possible answers reveals the respondent's true situation with certainty - even when he/she answers truthfully.

The main drawback of the method is that it provides the statistician with less information than a direct question would have done. At the end of the survey, the statistician's information is limited to the total numbers of answers 'correct' and 'false', which is less informative than the exact incidence of property A in the sample. This phenomenon reflects the statistician's and the respondent's conflicting interests: the former seeks a maximum of information, while the latter wants to protect his privacy by giving away as little information as possible. By a sensible choice of P a compromise may be reached. Warner called his method 'randomized response'; since its invention it has been enriched with a wealth of variations.

From his survey, the statistician can estimate the probability of occurrence of the answers 'correct' and 'false'. However, these probabilities are closely linked to the incidence of property A in the population. This can be shown as follows. Denote by π the population fraction showing property A, denote $\bar{A}$ for the absence of A and introduce θ for the probability of obtaining the answer 'correct' from a randomly chosen respondent. Then, assuming an honest answer,

$$\begin{aligned}\theta &= P(\text{'correct'}\,|\,A)P(A) + P(\text{'correct'}\,|\,\bar{A})P(\bar{A}) \\ &= P\pi + (1-P)(1-\pi) \\ &= 1 - P + (2P-1)\pi\end{aligned}$$

Hence, since P is a known constant, estimating π is essentially equivalent to estimating θ.

Being a population fraction, π has a value inside the unit interval. But then the above equations imply

$$1 - P \leq \theta \leq P$$

If the survey was based upon a simple random sample, the problem is essentially to estimate the parameter θ of a binomial distribution, where (unless $P = 1$) θ is known beforehand to lie within an interval strictly smaller than the natural parameter space $[0,1]$. This is an example of a so-called truncated parameter space, the object of study in this thesis.

The study of truncated parameter spaces in general is of interest for the following reasons. First of all, they often occur in practice. In many cases certain parameter values can be excluded beforehand, sometimes on strictly logical grounds, in other instances for theoretical reasons stemming from the nature of the problem under consideration; finally, the investigator can exclude certain parameter values for practical or subjective reasons. This last situation is rather frequent. Note however that the example of randomized response, treated above, strictly logically leads to a truncated parameter space, without any subjective arguments brought into the discussion. So, even strict frequentists should be interested in the subject.

Besides, truncated parameter spaces are mathematically and statistically interesting. Some of their most intriguing features, occurring under fairly general circumstances, are listed below.

(i) The frequently used criterion of unbiasedness is useless, since no unbiased estimators exist in general. So, other optimality criteria have to be looked for. As a leading optimality principle, admissibility has been adopted here; additional criteria will be invariance and minimaxity, among others.

(ii) Many common sense estimators that take values on the boundary of the truncated parameter space, or even close to it, are inadmissible. This means that no estimate should ever be near the extreme values of the parameter, even if all observations indicate such value as most likely. An important consequence is that maximum likelihood estimators are inadmissible in general.

(iii) The behavior of minimax estimators is much more complicated than in the classical, nontruncated case (as well as their derivation, for that matter). Furthermore, the behavior is sometimes rather surprising at first sight. For example, the minimax estimate corresponding to a certain observation that indicates a high parameter value may <u>de</u>crease, if the theoretical upperbound for the parameter <u>in</u>creases.

The organization of the book is as follows. Part 1 is devoted to the general theory of estimation in truncated parameter spaces, while the special case of estimating the parameter θ of a binomial distribution $B(n,\theta)$ is treated in Part 2. This includes the randomized response situation, outlined above, as a specific application. As a prelude to Part 2, many examples in Part 1 already deal with the binomial distribution. In many respects the treatment is based upon MOORS 1977.

Chapter 1 defines the kind of estimation problems to be discussed in the sequel, with the choice of the action space as the most striking feature: it will be recommended here to take as standard action space the convex closure of $\{h(\theta) : \theta \in \Theta\}$, where h denotes the function of the unknown parameter θ that has to be estimated. Attempts have been made to concentrate in Section 1.4 all standard estimation theory, that will be needed later on. The final Section 1.5 reviews some well-known properties of exponential families.

Chapter 2 focuses on truncated parameter spaces. After the definition in Section 2.1, the interesting features (i)-(iii), mentioned above, are illustrated for some special cases. Section 2.2 shows the non-existence of unbiased estimators in the case of exponential families and quadratic loss. In Section 2.3 estimators are considered with values upon or close to the boundary of the parameter space; for shortness, estimators of this kind will informally be called 'boundary rules'. The first examples are presented of truncated parameter spaces, where common sense boundary rules are inadmissible. In the final Section 2.4 some well-known theorems on minimax and Bayes estimators are applied to the truncated case. Among the examples is the case of the normal distribution as considered by CASELLA & STRAWDERMAN 1981 and BICKEL 1981.

One of the main results is described in Chapter 3, drawing heavily on MOORS 1981b. Here, invariance is accepted as additional optimality criterion, enabling the formulation of general results on the inadmissibility of boundary rules. The main Theorem 3.9 determines a strict subspace of the action space to which all estimates are restricted, conditionally on the observations. In this way, the class of potential estimators is reduced, excluding in general the maximum likelihood estimator, among others. From the numerous examples given the application to inequality constraint regression is believed to be the most interesting.

Chapter 4 concludes the general Part 1. It is devoted to exponential families, already introduced in Section 1.5. The results of the previous Chapter are applied to one-parametric exponential families in Section 4.2, while a similar analysis for two-parametric exponential families is started. In the remainder of the chapter Bayes estimators are considered in more detail. Attention is concentrated on the estimator derived by KATZ 1961 for a specific function $h(\theta)$; it is in fact a generalized Bayes rule with respect to the uniform measure. A detailed application can be found in Section 4.4, where a collection of admissible rules is derived for the expectation and variance of an exponential distribution. A fundamental quantity in this analysis is Mills' ratio for which the bounds, given by FELLER 1950, were simply proved and improved.

In Part 2 the special estimation problem is considered, featuring a binomial distribution $B(n,\theta)$ with unknown parameter θ, the truncated case of which was shown earlier to rise from Warner's randomized response model. Therefore, a more detailed account of the theory of randomized response is presented in Section 5.2; an overview is given of the most important developments since the review paper of HORVÍTZ et al. 1976. The next section deals with admissibility; here, Chapter 3 is applied to the binomial case. Section 5.4 considers Bayes estimators; it is shown that a Bayes rule only depends on the corresponding prior distribution through a limited number of its moments. Since a minimax estimator is Bayes with respect to a least favorable prior, the search for minimax estimators naturally leads to the question how to describe the moment space, that is the set of all vectors with components that equal exactly the first r moments of some probability distribution.

This celebrated moment problem is discussed in Chapter 6. Following KARLIN & SHAPLEY 1953 and KARLIN & STUDDEN 1966, a characterization of moment vectors is given by means of so-called Hankel determinants. A closely related question is how to find a probability distribution corresponding with a given moment vector. It is shown that in all situations discrete distributions exist; the construction method of VON MISES 1964 is used to find distributions with as few steps as possible.

The final Chapter 7 returns to the problem of minimax estimation. It can be solved now by finding a prior distribution that maximizes the minimum Bayes risk. This prior may be taken to be a discrete distribution with the minimum number of steps as indicated in Chapter 6. The minimax problem is thus reduced to a constrained optimization problem. The solution is based on the general algorithm of POWELL 1978; it is accounted for in detail in Section 7.3, that also presents the minimax estimators obtained. Section 7.2 describes the analytical solutions for very small sample sizes. The quadratic loss function, used up to here, is generalized in Section 7.4 to a weighted version; minimax estimators for this case are obtained as well.

A final word on notation. To avoid unnecessary brackets, the following convention is adopted. If $u : A \to B$ and $v : B \to C$ are arbitrary functions, the compound function $v \circ u : A \to C$ will be denoted by vu; hence, for any $a \in A$, $vu(a)$ indicates the element $(v \circ u)(a) = v(u(a))$ of C. For the sake of easy retrieval, within each chapter lemmas, figures, examples and other exhibits are numbered consecutively without distinction. The end of an example, definition or theorem is indicated by the symbol $\square$; if a theorem is proved, this symbol appears at the end of the proof.

PART 1

GENERAL THEORY

1. ESTIMATION THEORY

1.1. Introduction and summary

A very readable and lucid textbook on statistical decision theory is FERGUSON 1967; his approach and notation will to a great extent be followed here. The discussion starts in Section 1.2 with a general outline of the standard statistical decision problem and some important notions involved, like admissibility. In Section 1.3 a definition is given of regular estimation problems, to be considered in the sequel. An important feature of this definition is the space of all possible actions; a sensible choice of this space will appear to be the convex closure of the space of all possible values of the estimand. Furthermore, the loss function will be assumed to satisfy certain conditions, like convexity. It will be shown that these assumptions imply the essential completeness of the class of nonrandomized estimators.

Additional selection criteria for estimators, such as unbiasedness and invariance, are introduced and discussed in Section 1.4. Furthermore, minimax and Bayes rules are defined and discussed. Some theorems, more or less standard, are presented that will be needed in the sequel. In view of the considerations mentioned above, most results will be presented for nonrandomized estimators only, although many of them can readily be extended to more general decision rules.

The final Section 1.5 introduces exponential families of probability distributions. Some of their most interesting properties are reviewed.

1.2. General decision problems

Decision problems in general have to do with decision making under uncertainty. The effects of the decision depend on the true 'state of nature', which is not known at the time the decision has to be made. So, only ex post (at the best) it can be concluded to what extent a certain decision was correct. The dependence on the true state of affairs can be expressed more formally by stating that the effects of the decision are influenced by some parameter θ, the value of which is not known by the decision maker. It is of importance to try to obtain knowledge about θ, since better decisions may be reached if more information on θ is available.

In many cases information on θ can be obtained by observing a random variable X that has a probability distribution P_θ depending on θ. Distribution P_θ assigns probability $P_\theta\{X \in B\}$ to any Borel set B of the sample space X. Such decision problems are called statistical decision problems; a usual notation is (Θ, A, L, X). Here, Θ is the (given) parameter space, that is the space of all possible values of the parameter θ; A is the action space: the space of all actions (decisions) a, that are theoretically possible or which the decision maker is prepared to consider. Throughout the book both Θ and A will be assumed to be Borel sets of finite dimensional Euclidean spaces, to be denoted by $\mathbb{R}_k$ and $\mathbb{R}_m$, respectively. The loss function $L : \Theta \times A \to \mathbb{R}$ denotes the loss suffered, when action $a \in A$ was taken and $\theta \in \Theta$ is the true value of the parameter; $L(\theta,a)$ is assumed to be Borel measurable in the pair (θ,a). If possible, the best solution of course would be to choose for all possible observations $x \in X$ an action $a \in A$ that minimizes the loss, whatever the true value of θ. Unfortunately, decisions for which the loss is minimum, uniformly in θ, only exist in trivial situations.

It may be useful to extend space A of possible actions in the following way: leave the final choice of a decision to a chance mechanism that selects an action $a \in A$ according to a prescribed probability distribution. This line of reasoning leads to the introduction of a larger action space A^*, consisting of probability distributions a^* on the Borel sets of A. Rather than making a decision $a \in A$ directly, the decision maker selects a so-called randomized decision $a^* \in A^*$ and plugs in the chance mechanism to pick the final action (with probabilities determined by a^*). Let Z denote a random variable on A with distribution a^* and let EZ denote its expectation with respect to a^*. The domain of L can be extended to $\Theta \times A^*$ by defining the loss corresponding with a^* as $EL(\theta,Z)$, provided this expectation exists and is finite for all $\theta \in \Theta$. In other words, A^* is defined as the space of all probability distributions a^* on A with finite loss for all $\theta \in \Theta$. By identifying degenerate distributions from A^* with the corresponding actions in A, the latter can be formally identified with a subspace of A^*. With these definitions problem (Θ, A, L, X) can be extended to the richer decision problem (Θ, A^*, L, X) in which randomization is allowed.

Statistical decision problems should be analyzed before the observations have been taken into account. Consequently, this analysis must deal with all possible realisations x of X. This leads to the study of decision rules. The most general ones are behavioral decision rules: measurable mappings $\delta: X \to A^*$, prescribing for any $x \in X$ which action $a^* \in A^*$ must be taken. The average loss suffered from frequent application of δ is given by $E_\theta L(\theta, \delta(X))$, where E_θ denotes the expectation with respect to probability distribution P_θ. All rules δ to be considered will have the property that this expectation is finite for all $\theta \in \Theta$. The class of all decision rules δ with finite expected loss will be denoted by D.

The statistical decision problem can now be reconsidered as the question how to select the 'best' decision rule $\delta \in D$, in a sense to be made more precise. A first important feature of a rule δ is its expected loss; it is called the risk function $R : \Theta \times D \to \mathbb{R}$, defined by $R(\theta,\delta) := E_\theta L(\theta, \delta(X))$. The lower the risk the better of course. Unfortunately again, rules for which the risk function is minimum uniformly in θ only exist in trivial cases. It is more rewarding to compare the risk functions of two different decision rules: if either of them is uniformly higher, it stands to reason that the corresponding rule is worse and may safely be discarded. More precisely: rule $\delta_1 \in D$ is said to dominate $\delta_2 \in D$ if $R(\theta,\delta_1) \leq R(\theta,\delta_2)$ holds for all $\theta \in \Theta$, and to dominate strictly if, moreover, $R(\theta,\delta_1) < R(\theta,\delta_2)$ holds for some $\theta \in \Theta$. If $R(\theta,\delta_1) = R(\theta,\delta_2)$ holds for all $\theta \in \Theta$, δ_1 and δ_2 are called equivalent. Now, $\delta_1 \in D$ is called admissible if no $\delta \in D$ dominates it strictly. Domination induces a partial ordering in D; by restricting attention to admissible rules, D is cleared of all rules with a uniformly higher (or equal) risk function.

In this sense, admissibility is a first and rather fundamental optimality criterion. It is applicable in all situations and constitutes an almost undisputed central property. The only exception sometimes arises in practical cases, where inadmissible but quick-and-easy rules may be preferred to admissible but cumbersome ones. Compare TUKEY 1977.

Class C of decision rules from $\mathcal{D}$ is called essentially complete, if any $\delta \in \mathcal{D}$ is dominated by some $\delta' \in C$. As a consequence of the central part of the notion of admissibility, attention may be restricted to essentially complete classes of decision rules. (Note however, that equivalent rules may be discarded in this way.) An example of such a class constitute the decision rules based upon a sufficient statistic; see FERGUSON 1967, Th. 3.4.1. Intuitively, this statement is immediately clear. Recall that a statistic $T := t(X)$ with t a Borel function on $\mathcal{X}$, is sufficient for θ, $\theta \in \Theta$, if the conditional distribution of X, given $\{T = t\}$, does not depend on θ for all $t \in \mathcal{T} \subset t(\mathcal{X})$ with $P_\theta\{T \in \mathcal{T}\} = 1$ for all $\theta \in \Theta$. More informally speaking, a sufficient statistic T carries all information on θ that is stored in the variable X. Hence, reduction of $\mathcal{D}$ by sufficiency cannot result in the exclusion of important decision rules and will be applied in the sequel where possible.

A decision rule that maps $\mathcal{X}$ into $\mathcal{A}$ rather than $\mathcal{A}^*$ is called nonrandomized and will be denoted by d. Since d directly prescribes what action is to be taken without interference of the chance mechanism, a nonrandomized rule has the advantage of greater simplicity. Besides, the class D of all nonrandomized rules from $\mathcal{D}$ is essentially complete under fairly general conditions. Precise conditions are given in the next section; since they will appear to be satisfied in the decision problems to be met in the sequel, nonrandomized rules will occur almost exclusively from now on.

Since admissibility only leads to a partial, not a complete, ordering of D, additional optimality concepts are needed. The classical literature presents four such concepts, to be subdivided into two categories. The first category contains optimality criteria defining a certain subset of D; other rules are simply not under consideration anymore. The following two subsets are generally considered: the set of

(i) unbiased rules;

(ii) invariant rules.

The second category contains two additional optimality criteria; each of them induces a complete ordering in D and therefore leads to an optimal decision rule:

(iii) minimax criterion;

(iv) Bayes criterion.

These four concepts will be defined and discussed in Section 1.4.

1.3. Regular estimation problems

Traditionally, statistical decision theory is subdivided into hypothesis testing and estimation theory. In testing problems, the central question is whether θ belongs to a given subspace Θ_0 of Θ, or rather to its complement; hence A typically consists of exactly two points. Estimation problems concern the choice of an assumed value for the estimand, which should be as close as possible to the true value. In this case, A typically contains an open set in $\mathbb{R}_m$. Since this thesis is about estimation, only the latter situation will be considered. Therefore, the word 'decision rule' will be avoided in the sequel; instead, 'estimation' or 'rule' will be used.

Not necessarily the parameter θ has to be estimated; more generally, the estimand is a given Borel function $h : \Theta \to \mathbb{R}_m$. This situation will be adopted as standard here. The loss function will depend on a only through $h(\theta)-a$; hence

$$(1.1) \qquad L(\theta,a) = L^*(\theta,h(\theta)-a)$$

for some real-valued L^* with appropriate domain in $\Theta \times \mathbb{R}_m$. Typically, L is increasing in $|h(\theta)-a|$ and often also convex in a for all $\theta \in \Theta$. Common loss functions are the <u>quadratic loss function</u> $|h(\theta)-a|^2$ and the <u>absolute loss function</u> $|h(\theta)-a|$; both have the properties mentioned above. A more sophisticated example of a type (1.1) loss function is $w(\theta)|h(\theta)-a|^2$, which will sometimes be used as well.

For some estimation problems the set of all possible <u>estimates</u> (actions) is unambiguously determined by the nature of the problem itself and cannot be influenced by the statistician. In the majority of cases, however, the statistician can freely choose beforehand which actions he is willing to consider at all. His main concern will be not to exclude a priori any possibly valuable actions; so the action space should be large enough. It follows as a logical demand that all possible values of $h(\theta)$ can indeed be obtained as estimates; this implies $A \supset h(\Theta)$, where the abbreviation $h(\Theta) := \{h(\theta) : \theta \in \Theta\}$ has been used. The choice $A = h(\Theta)$ is usual in estimation problems; see, for example, FERGUSON 1967, p. 11 or DE GROOT 1970, p. 226. However, in this study a larger action space is chosen, namely the convex closure of $h(\Theta)$. (Recall that the convex closure $C\{S\}$ of some $S \subset \mathbb{R}_n$ is the intersection of all closed convex sets in $\mathbb{R}_n$ that contain S.) This choice will be discussed now.

First of all, action space $A = C\{h(\Theta)\}$ is large enough in the sense that no possibly useful actions are excluded. To see this, take any $a \notin C\{h(\Theta)\}$ and let a_0 denote its perpendicular projection on (the closed space) $C\{h(\Theta)\}$. Then convexity ensures that $|h(\theta)-a| > |h(\theta)-a_0|$ holds for all $\theta \in \Theta$; compare Figure 1.1. If $L(\theta,a)$ is increasing in $|h(\theta)-a|$, this implies $L(\theta,a) > L(\theta,a_0)$ for all $\theta \in \Theta$; in other words, a_0 induces a loss that is smaller than the loss from a, uniformly in θ. So $a \in A$ is never preferred to $a_0 \in A$.

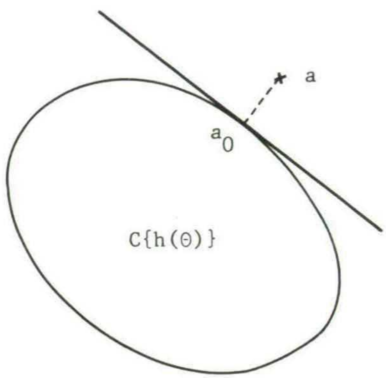


Figure 1.1. The (convex) action space $C\{h(\Theta)\}$

The fact that the recommended action space is closed implies the desirable property that the limit of a convergent series of actions in A belongs to A as well. For instance, consider the class of nondegenerate binomial distributions $B(n,\theta)$, which has the open unit interval as (natural) parameter space. In estimating θ, it is convenient to have the estimates 0 and 1 at one's disposal as well, leading to the closed unit interval as action space.

The main argument for taking the convex closure of $h(\Theta)$ as action space throughout this thesis, is mathematical convenience: class D of nonrandomized rules becomes essentially complete under rather mild conditions (FERGUSON 1967, Th. 2.8.1).

Theorem 1.2. Let A be convex and $L(\theta,a)$ convex in a for all $\theta \in \Theta$; assume that for some $\theta \in \Theta$ an $\varepsilon > 0$ and a $c \in \mathbb{R}$ exist, such that $L(\theta,a) \geq \varepsilon|a| + c$ for all $a \in A$. Then, for any $a^* \in A^*$, an $a \in A$ exists such that $L(\theta,a) \leq L(\theta,a^*)$ for all $\theta \in \Theta$.

Proof. Let Z denote a random variable on A with distribution a^*. Then $\varepsilon EZ + c \leq EL(\theta,Z) = L(\theta,a^*) < \infty$, so that $E(Z) < \infty$. Now choose $a = E(Z)$. Then Jensen's inequality gives

$$L(\theta,a^*) = EL(\theta,Z) \geq L(\theta,EZ) = L(\theta,a)$$

for all $\theta \in \Theta$. □

So, under the conditions of the theorem, for any randomized action a nonrandomized action can be found with a loss that is not higher - uniformly in θ. Extension of the theorem to estimators shows that D is essentially complete under the conditions given. Since the quadratic as well as the absolute loss function both satisfy these conditions, this result is of great practical importance. Note that the somewhat peculiar condition $L(\theta,a) \geq \varepsilon|a| + c$ is trivially satisfied, if A is bounded; for unbounded A it is equivalent to the condition that a $\theta \in \Theta$ exists for which $L(\theta,a) \to \infty$ if $|a| \to \infty$ (FERGUSON 1967, Exercise 2.8.3).

The following property may be considered a marked disadvantage of the action space $C\{h(\Theta)\}$: this choice allows estimates that fall in a subset of A, which with certainty does not contain the true value of the estimand. The seriousness of this disadvantage should not be exaggerated, however. If the loss function was correctly chosen, admissibility should be accepted as the leading optimality property. Otherwise the loss function should penalize more severely these 'impossible' estimates. The following example presents an illustration.

Example 1.3. Consider the estimation problem (Θ,A,L,X) with $L(\theta,a) = (\theta-a)^2$, where X has the binomial distribution $B(2,\theta)$. (An easy interpretation is that the probability θ of throwing heads has to be estimated from two tosses with a certain coin.) If Θ and A both are the open unit interval (0,1),

$$d_m(0) = c, \quad d_m(1) = \tfrac{1}{2}, \quad d_m(2) = 1 - c$$

where $c := (\sqrt{2}-1)/2$, defines an estimator d_m for θ; it is easy to show that d_m has constant risk c^2. However, take $\Theta = (0,1-P] \cup [P-1)$ for some $P \in \mathbb{R}$ with $\frac{1}{2} < P < 1$. (A possible interpretation is that the coin is known to have a bias of at least $P - \frac{1}{2}$, in an unknown direction.)

First, take $A = \Theta$ as usual. Consider any rule $d \in D$ for this problem and write $x := d(0)$, $y := d(1)$ and $z := d(2)$ for convenience. The risk equals

$$R(\theta,d) = (1-\theta)^2(x-\theta)^2 + 2\theta(1-\theta)(y-\theta)^2 + \theta^2(z-\theta)^2$$

and has the property

$$R(\theta,d) > c^2 \text{ for some } \theta \in \Theta \tag{1.2}$$

To prove this, assume that $R(0,d) \leqslant c^2$ and $R(1,d) \leqslant c^2$, giving $x \leqslant c$ and $z \geqslant 1-c$ respectively. Then

$$\max\{R(P,d),R(1-P,d)\} - c^2$$

$$\geqslant (c-P)^2(1-P)^2 + (1-2P)^2 2P(1-P) + (1-c-P)^2P^2 - c^2$$

$$= 2P(1-P)[P(1-P) + (1-2P)^2 - c(c+1)]$$

$$= 6P(1-P)(P-\tfrac{1}{2})^2 > 0$$

so that (1.2) holds for $\theta = P$ or $\theta = 1-P$.

Now, take $A = C\{\Theta\}$. Then rule d_m again has constant risk c^2 and none of the alternative rules, considered in the previous case, dominates it. (It can be shown similarly that d_m is not dominated either by the behavioral rules δ, where $\delta(1)$ takes the values P and 1-P with probability $\frac{1}{2}$ each.) In short, the choice $A = \Theta$ excludes the (minimax) rule d_m, which would be admissible with the choice $A = C\{\Theta\}$. □

In view of the above arguments, the following definitions will be adopted.

Definition 1.4. An estimation problem consists of the following elements:

(i) a given parameter space $\Theta \subset \mathbb{R}_k$;

(ii) a given function $h : \Theta \to \mathbb{R}_m$;

(iii) the action space $A = C\{h(\Theta)\} \subset \mathbb{R}_m$;

(iv) a given loss function $L : \Theta \times A \to \mathbb{R}$;

(v) a random observable X with some distribution P_θ, $\theta \in \Theta$.

Further, L is measurable in the pair (θ, a), allowing a representation

$$(1.1) \qquad L(\theta,a) = L^*(\theta,h(\theta)-a)$$

and $h(\Theta)$ contains an open set in $\mathbb{R}_m$. □

Note that the domain of L^* is $\Theta \times U$, where U consists of all vectors $u \in \mathbb{R}_m$, that can be written as the difference of two vectors, in $h(\Theta)$ and $C\{h(\Theta)\}$ respectively.

Definition 1.5. The estimation problem of Definition 1.4 is called regular, if $L(\theta,a)$ has the following properties:

(i) L is convex in a for all $\theta \in \Theta$;

(ii) if A is unbounded, then a $\theta \in \Theta$ exists for which $L(\theta,a) \to \infty$ as $|a| \to \infty$.

In the special case of quadratic loss function

$$(1.3) \qquad L(\theta,a) = |h(\theta) - a|^2$$

the estimation problem is called quadratic. □

Estimation problems will be denoted by (Θ,L,X) in the sequel. Contrary to the usual notation of decision problems, A has been suppressed, since it is fully determined by the other entities. Likewise, h has been suppressed in the notation, as it is implicitly determined by the loss function.

Corollary 1.6. For a regular estimation problem class D of nonrandomized estimators is essentially complete.

Proof. This statement is an immediate consequence of Theorem 1.2: A and L are convex by definition, and the additional condition on L was seen to be equivalent with property (ii) in Definition 1.5. □

In typical estimation problems, $L^*(\theta,u)$ is increasing in u ; in combination with convexity this ensures property (ii) in Definition 1.5. In the sequel attention will be concentrated on regular estimation problems, hence on nonrandomized estimators. For that reason all concepts in the next section are formally defined only for nonrandomized rules, although more general definitions are possible.

1.4. Regular estimation theory

The four additional optimality criteria, mentioned at the end of Section 1.2, will be defined and discussed now. The definitions given are not the most general possible, but specialized to (regular) estimation problems; in particular, they apply to nonrandomized estimators.

(i) The well-known concept of unbiasedness is defined as follows for estimators: $d \in D$ is called unbiased for $h(\theta)$ if its expectation exists and if $E_\theta d(X) = h(\theta)$ holds for all $\theta \in \Theta$. Note that this definition does not depend on the choice of the loss function, contrary to Lehmann's more general definition; see LEHMANN 1959, Section 1.5. Recently, NOORBALOOCHI & MEEDEN 1983 gave an even more general definition of unbiasedness, based on Bayesian ideas.

(ii) The invariance principle is applicable only to problems showing some kind of symmetry. Consider a group G of Borel functions g of $\mathcal{X}$ onto itself. The class of distributions $\{P_\theta : \theta \in \Theta\}$ is called invariant (or symmetric) under G, if of the following conditions (a) is satisfied. If condition (b) holds as well, estimation problem (Θ,L,X) is called invariant under G.

(a) For every $g \in G$ and every $\theta \in \Theta$ a unique $\bar{g}(\theta) \in \Theta$ exists, such that the distribution of $g(X)$ is given by $P_{\bar{g}(\theta)}$, whenever P_θ is the distribution of X. This implies

$$(1.4) \qquad P_{\bar{g}(\theta)}\{X \in B\} = P_\theta\{X \in g^{-1}(B)\}$$

for all Borel sets $B \subset X$. Note that g^{-1} exists and belongs to G, since G is a group.

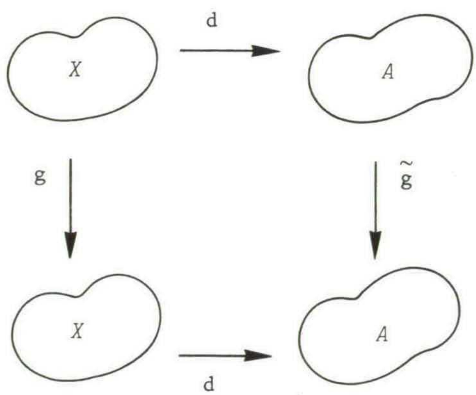
(b) For every $g \in G$ and every $a \in A$ a unique $\tilde{g}(a) \in A$ exists such that

$$L(\bar{g}(\theta),\tilde{g}(a)) = L(\theta,a), \quad \theta \in \Theta \tag{1.5}$$

Note that in an invariant estimation problem every $g \in G$ induces two surjections, namely $\bar{g} : \Theta \to \Theta$ and $\tilde{g} : A \to A$; the sets $\bar{G} := \{\bar{g} : g \in G\}$ and $\tilde{G} := \{\tilde{g} : g \in G\}$ are groups again. Now, for an invariant problem, the nonrandomized estimator $d \in D$ is called <u>invariant</u> (or <u>equivariant</u>) <u>under G</u> if

$$dg(x) = \tilde{g}d(x) \quad , \; x \in X$$

holds for every $g \in G$. An intuitive explanation of this relation reads as follows. Simultaneous application of the transformations g, $\bar{g}$ and $\tilde{g}$ does not change the relations between the elements of X, Θ and A respectively, nor the structure of the problem (Θ,L,X). Hence, this operation can be viewed as a simple relabeling of the elements of X, A (and Θ), that should not affect the estimator in any way. Figure 1.7 illustrates the reasoning.



<u>Figure 1.7</u>. Schematic picture of an invariant estimator d

An important property, which will be needed in the sequel, is that

(1.6) $\quad R(\theta,d) = R(\bar{g}(\theta),d),\ \theta \in \Theta$

holds for every invariant $d \in D$ and every $\bar{g} \in \bar{G}$.

(iii) The minimax principle induces a complete ordering in D. A ruled $d_m \in D$ is called <u>minimax</u> if

(1.7) $$\sup_{\theta \in \Theta} R(\theta,d_m) = \inf_{d \in D} \sup_{\theta \in \Theta} R(\theta,d)\ (\leq \infty)$$

holds. According to this principle the statistician mitigates the worst that may happen; it corresponds with a risk avoiding behaviour.

(iv) The Bayesian view considers the true parameter value θ not as deterministic, but as the realisation of a random variable having some probability distribution, called the <u>prior distribution</u> τ. The <u>Bayes risk</u> of estimator $d \in D$ with respect to τ is defined by $r(\tau,d) := ER(Z,d)$, where Z is a random variable on Θ with distribution τ. Now, rule $d_\tau \in D$ is called <u>Bayes with respect to τ</u>, if

(1.8) $$r(\tau,d_\tau) = \inf_{d \in D} r(\tau,d)$$

So, contrary to minimax rules that minimize the maximum risk, Bayes rules minimize the average risk (with the prior distribution as weighting function).

Different Bayesian statisticians will have different prior distributions on a given θ. The class of all prior distributions τ is denoted by Θ^*; then the Bayes risk is defined by the function $r : \Theta^* \times D \to \mathbb{R}$. A prior distribution τ_* is <u>least favorable</u>, if

(1.9) $$\inf_{d \in D} r(\tau_*,d) = \sup_{\tau \in \Theta^*} \inf_{d \in D} r(\tau,d)$$

Hence, if the parameter value θ were chosen according to τ_*, the minimum Bayes risk $r(\tau,d_\tau)$ would be maximized.

The case that τ is not a proper probability distribution, but merely a σ-finite measure on Θ, is of interest as well. A rule $d \in D$ mimimizing $r(\tau,d) := \int R(\theta,d)\tau\{d\theta\}$ is called generalized Bayes with respect to the measure τ. Note that the existence of a (generalized) Bayes rule is not guaranteed, nor is its uniqueness.

In practical applications in particular, the optimality concepts discussed above will sometimes be overruled by considerations of simplicity and tractability: if an admissible, but complicated rule only slightly dominates a quick-and-easy rule, the latter may be preferred. Furthermore, the different criteria may be combined in many ways. E.g., a minimax estimator can be defined within the class of invariant estimators; the famous BLU estimator combines property (i) with linearity - which can be viewed as an elaboration of the simplicity consideration. Another example is the concept of T-minimaxity, an attempt to combine the properties (iii) and (iv), which was introduced by HODGES & LEHMANN 1952. Assume that the statistician is able to specify a subset $T \subset \Theta^*$, to which the prior distribution must belong. Then rule $d \in D$ is called T-minimax if

$$\sup_{\tau \in T} r(\tau,d) = \inf_{d' \in D} \sup_{\tau \in T} r(\tau,d')$$

Note that a T-minimax rule reduces to a Bayes rule for $T = \{\tau\}$ and to a minimax rule for $T = \Theta^*$. Applications to binomial estimation were presented by BLUM & ROSENBLATT 1967 and JACKSON et al. 1970.

The remainder of this section presents some important theorems, that will be needed in the sequel. Where reference is possible, no formal proof will be given, but only some explanatory notes. Although results will be needed for regular estimation problems only, many of the following theorems hold in more general situations as will be shown by their formulation.

To start with, two theorems on admissibility of Bayes rules are presented. The first demands uniqueness of the Bayes rule involved.

Theorem 1.8. If a (generalized) Bayes rule with respect to a given prior τ is unique up to equivalence, this estimator is admissible. $\square$

See FERGUSON 1967, Theorem 2.3.1. The proof runs along the following lines. Let d_τ be (generalized) Bayes with respect to τ; assume that a $d \in D$ exists dominating d_τ, so that $R(\theta,d) \leqslant R(\theta,d_\tau)$ for all $\theta \in \Theta$. Then

$$r(\tau,d) \leqslant r(\tau,d_\tau) = \inf_{d' \in D} r(\tau,d')$$

so that $r(\tau,d) = r(\tau,d_\tau)$ and d is (generalized) Bayes with respect to τ as well. The uniqueness ensures that d is equivalent to d_τ.

The second theorem on admissibility of Bayes rules requires continuity of the risk function as a major condition. Since it concerns a slight generalization of Theorem 2.3.3 in FERGUSON 1967, a complete proof is given. Recall that the support of a distribution $\tau \in \Theta^*$ is defined as the set of all $\theta_0 \in \Theta$ such that for all $\varepsilon > 0$ ball $\{\theta \in \Theta : |\theta-\theta_0| < \varepsilon\}$ has positive probability.

Theorem 1.9. Let Θ be open and assume that $R(\theta,d)$ is a continuous function of θ for every $d \in D$. If d_τ is (generalized) Bayes with respect to a given prior τ with support Θ, while $r(\tau,d_\tau)$ is finite, then d_τ is admissible.

Proof. Assume that d_τ is dominated strictly by some $d' \in D$. Then there exists a $\theta_0 \in \Theta$ such that $\eta := R(\theta_0,d_\tau) - R(\theta_0,d') > 0$. From the continuity of R follows the existence of an $\varepsilon > 0$ for which $R(\theta,d') \leqslant R(\theta,d_\tau) - \eta/2$, whenever $|\theta-\theta_0| < \varepsilon$. Hence,

$$\int [R(\theta,d') - R(\theta,d_\tau)]\tau\{d\theta\} \geqslant \frac{\eta}{2}\,\tau\{\theta : |\theta-\theta_0| < \varepsilon\}$$

holds; the right-hand side is positive, since θ_0 is in the support of τ. This contradicts the fact that d_τ is (generalized) Bayes with respect to τ. □

Sufficient conditions for the risk function to be continuous in the above sense will be given in the next section.

A central quantity in the definition of a least favorable prior proved to be $\sup_\tau \inf_d r(\tau,d)$, the right-hand side of (1.9) in a condensed notation. Similarly, (1.7) shows that for a minimax rule $\inf_d \sup_\theta R(\theta,d)$

is of importance, which is easily seen to be equal to $\inf_d \sup_\tau r(\tau,d)$. Now, the inequality

$$\inf_{d' \in D} r(\tau,d') \leq \sup_{\tau' \in \Theta^*} r(\tau',d)$$

is true for all $\tau \in \Theta^*$ and all $d \in D$, of course. Taking $\sup_\tau$ on the left and $\inf_d$ on the right gives the general relation

$$\text{(1.10)} \qquad \sup_{\tau \in \Theta^*} \inf_{d \in D} r(\tau,d) \leq \inf_{d \in D} \sup_{\tau \in \Theta^*} r(\tau,d)$$

If equality holds here, the common value is called the <u>(minimax) value</u> of the problem. The next theorem gives conditions for an estimation problem to have a value. Recall that a function $f : S \to \mathbb{R}$ on some topological space S is <u>lower semicontinuous</u>, if for all $c \in \mathbb{R}$ the set $\{s \in S : f(s) > c\}$ is open; of course, continuity implies lower semicontinuity.

<u>Theorem 1.10</u>. Let $C \subset D$ be an essentially complete class. Assume that a topology on C exists such that C is compact and $R(\theta,d)$ is lower semicontinuous in $d \in C$ for all $\theta \in \Theta$. Then the estimation problem has a value and a minimax rule exists. □

Since $\sup_\theta R(\theta,d)$ is lower semicontinuous and is defined on a compact set, it attains its infimum at some point $d_0 \in C$, which is the minimax rule. The proof that the problem has a value, is rather technical; see FERGUSON 1967, Theorem 2.9.2.

The discussion concerning inequality (1.10) suggests that for problems with a value, the Bayes rule with respect to the least favorable distribution is minimax as well. This appears to be true under fairly general conditions.

<u>Theorem 1.11</u>. Assume that a given estimation problem has a value and that a minimax rule exists as well as a least favorable prior τ_*. If d_0 is the unique Bayes rule with respect to τ_*, then d_0 is minimax.

Proof. Let d_1 be minimax, so that $\sup_\tau r(\tau,d_1) = \inf_d \sup_\tau r(\tau,d)$. Since the problem has a value, the right-hand side equals $\sup_\tau \inf_d r(\tau,d) = \inf_d r(\tau_*,d)$, implying that d_1 is Bayes with respect to τ_*. The uniqueness of d_0 completes the argument. □

The next result shows, that the search for minimax estimators can be confined to invariant rules; compare FERGUSON 1967, Theorem 4.3.1. Similarly, if a least favorable prior exists, there also exists an invariant one; here, an invariant distribution $\tau \in \Theta^*$ satisfies for any $\bar{g} \in \bar{G}$

$$(1.11) \qquad \tau\{\bar{g}^{-1}(B)\} = \tau\{B\}$$

for all Borel sets $B \subset \Theta$.

Theorem 1.12. Assume that estimation problem (Θ,L,X) is invariant under a finite group G. If an estimator is minimax within the class of invariant estimators, it is minimax. □

Theorem 1.13. Assume that estimation problem (Θ,L,X) is invariant under a finite group G. If a prior distribution is least favorable within the class of invariant prior distributions, it is least favorable.

Proof. Let $\tau \in \Theta^*$ be least favorable; define τ_0 by

$$\tau_0\{B\} := \frac{1}{m} \Sigma \tau\{\bar{g}(B)\}$$

for all Borel sets $B \subset \Theta$, where the summation applies to all m elements $\bar{g} \in \bar{G}$. Obviously, τ_0 is invariant, and least favorable by (1.9). □

As an application of the various concepts and theorems presented, a well-known example is treated concerning the estimation of the parameter of a binomial distribution.

Example 1.14. Consider the problem of estimating the probability θ of throwing heads with a given coin, where $0 < \theta < 1$. The observations $Y_1,\ldots,Y_n$ consist of the results (heads or tails) of n (independent) tosses with the coin; squared error loss is used. More formally stated, this is a quadratic estimation problem (Θ,L,Y) with $\theta = (0,1)$, $L(\theta,a) =$

$(\theta-a)^2$ and $Y = (Y_1,\ldots,Y_n)$ with $Y_i \in B(1,\theta)$ and all Y_i independent; h is the identity and $A = [0,1]$.

First of all, $X := \Sigma_i Y_i$, the number of heads in the series of n tosses, is a sufficient statistic for θ, $\theta \in \Theta$; so, the above estimation problem may be reduced to (Θ,L,X), where $X \in B(n,\theta)$. According to Corollary 1.6 class D of nonrandomized rules (based on X) is essentially complete. D can be identified with $[0,1]^{n+1}$, hence is compact, so that Theorem 1.10 guarantees the existence of a minimax rule - even a nonrandomized one, say d_m. It can be shown that d_m is given by

$$(1.12) \qquad d_m(x) = \frac{x+\sqrt{n}/2}{n+\sqrt{n}}, \quad x = 0,1,\ldots,n$$

with risk $(\sqrt{n}+1)^{-2}/4$, independent of θ. (Note that the rule d_m, defined in Example 1.3, satisfies (1.12) with $n = 2$ and is minimax, by consequence.) By virtue of Theorem 1.13 an invariant least favorable prior distribution exists; the beta distribution $Be(\sqrt{n}/2,\sqrt{n}/2)$ presents an example. Many other least favorable priors exist; e.g. for $n = 1$, any prior τ with $E_\tau\theta = \frac{1}{2}$ and $E_\tau\theta^2 = \frac{3}{8}$ is least favorable as well. By virtue of Theorem 1.9 d_m is admissible. See for a detailed discussion FERGUSON 1967, p. 93 ff, or the very appealing reasoning in STEINHAUS 1957.

Further, (Θ,L,X) is invariant under the group G consisting of the identity $e : X \to X$ and the function g defined by $g(x) = n-x$, $x \in X$. The corresponding functions $\bar{g}$ and $\tilde{g}$ are given by $\bar{g}(\theta) = 1-\theta$, $\theta \in \Theta$ and $\tilde{g}(a) = 1-a$, $a \in A$; invariant estimators have to satisfy $d(n-x) = 1-d(x)$. It follows that d_m is invariant; indeed, according to Theorem 1.12 an invariant minimax rule must exist.

Finally, it can be shown that if the (weighted quadratic) loss function $L(\theta,a) = (\theta-a)^2/\{\theta(1-\theta)\}$ were used, the sample fraction X/n would be minimax for θ (FERGUSON 1967, Exercise 2.11.8). □

1.5. Exponential families

In an estimation problem (Θ,L,X), the statistician must in fact select a subset of the class of probability distributions $\{P_\theta : \theta \in \Theta\}$, to which the true distribution will belong. It is therefore logical to study classes of distributions defined on the same sample space X and look for common properties.

When P_θ is absolutely continuous with respect to a common σ-finite measure μ for all $\theta \in \Theta$, the class $\{P_\theta : \theta \in \Theta\}$ is said to be dominated (by μ). The (Radon-Nikodym) derivative of P_θ with respect to μ is called the density and will be denoted by $f(x|\theta)$; hence

$$P_\theta\{X \in B\} = \int_B f(x|\theta)\mu\{dx\}$$

holds for all Borel sets $B \subset X$. The most important dominated classes arise, when μ is the counting measure (then all P_θ are discrete and f is in fact a point mass function) or when μ is the Lebesgue measure; all classes of distributions to be met in the examples will belong to one of these two types.

Within a dominated class smaller classes of probability distributions may be distinguished. Perhaps the best known examples are exponential families which have as main advantage that they allow an impressive reduction of the data. Besides, the most familiar probability distributions belong to an exponential family.

An exponential family is a dominated class of probability distributions $\{P_\theta : \theta \in \Theta\}$, all of them having a density of the form

$$(1.13)\qquad f(x|\theta) = \exp[\sum_{i=1}^{k} \pi_i(\theta)t_i(x) + a(x) - c(\theta)] \ , \ x \in X$$

where all functions, occurring in the right-hand side, are real-valued. If either the functions $1,\pi_1,\ldots,\pi_k$ on Θ or the functions $1,t_1,\ldots,t_k$ on X are linearly dependent, the number k can be reduced. Further, the analysis is simplified by considering $\pi := (\pi_1,\ldots,\pi_k)$ as parameter, rather than θ. The natural parameter space Π, corresponding with this so-called natural parameter π, is defined as the set of all $\pi \in \mathbb{R}_k$ satisfying

$$\int_X \exp[\sum_{i=1}^{k} \pi_i t_i(x) + ax]\mu\{dx\} < \infty$$

It is not difficult to show that Π is convex and contains an open set in $\mathbb{R}_k$; compare LEHMANN 1983, Section 1.4. It will be useful to assume that this natural parameter space is open. Hence, the following definitions are given.

Definition 1.15. A dominated class of probability distributions $\{P_\theta : \theta \in \Theta\}$ is called an exponential family, if for all $\theta \in \Theta$ the density of P_θ can be written as (1.13). □

Definition 1.16. An exponential family is called k-parametric, if the two following conditions are satisfied:

(i) the functions $1, \pi_1, \ldots, \pi_k$ are linearly independent on Θ;

(ii) the functions $1, t_1, \ldots, t_k$ are linearly independent on X outside any Borel set of μ-measure 0. □

Definition 1.17. A (k-parametric) exponential family is called regular, if $\pi(\Theta)$ is open and convex. □

Definition 1.18. A regular k-parametric exponential family has its canonical form, if $\pi(\theta) = \theta$ for all $\theta \in \Theta$. □

See BARNDORFF-NIELSEN 1978 for a detailed description of exponential families and their properties.

The most important properties which will be used in the sequel are the following. First of all, exponential families are closed under sampling, which means that a random sample from distribution (1.13) can be written in the same form. Further, it is obvious that the set on which densities (1.13) are positive, does not depend on θ; in other words, all densities $f(x|\theta)$ have the same support. For any regular k-parametric exponential family statistic $t(X) := (t_1(X), \ldots, t_k(X))$ is complete sufficient for $\theta \in \Theta$. Recall that a statistic $t(X)$ is called complete for $\theta \in \Theta$, if the following statement is true: $E_\theta t(X) = 0$ for all $\theta \in \Theta$ implies $t(x) = 0$ a.e. (with respect to all P_θ). For regular families in the canonical form $c(\theta)$ is infinitely often differentiable for all $\theta \in \Theta$; the expectation and the covariance matrix of $t(X)$ satisfy for all $\theta \in \Theta$:

$$(1.14) \qquad E_\theta t(X) = \nabla c(\theta)$$

(1.15) $\quad V_\theta t(X) = Hc(\theta) > 0$

respectively, where ∇ and H denote gradient and Hessian. These identities can easily be found by differentiating equality $\int f(x|\theta)\mu\{dx\} = 1$; the positivity in (1.15) follows from the linear independence of the components of t(X). Finally, note that notation (1.13) for k-parametric exponential families is not unique, since linear transformations of the functions on the right-hand side are possible which leave the exponent unchanged.

<u>Example 1.19</u>. The class of binomial distributions $\{B(n,\theta) : 0 < \theta < 1\}$ is a one-parametric exponential family with density (with respect to the counting measure)

$$f^*(x|\theta) = \exp[x \log \frac{\theta}{1-\theta} + \log\binom{n}{x} + n \log(1-\theta)] \ , \ x = 0,1,\ldots,n$$

Introduction of the natural parameter $\pi := \log[\theta/(1-\theta)]$ leads to the regular canonical form

(1.16) $$f(x|\pi) = \exp[\pi x + \log\binom{n}{x} - n \log(1+e^\pi)] \ , \ x = 0,1,\ldots,n$$

The normal distributions $\{N(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ constitute a regular two-parametric exponential family with

(1.17) $$f(x|\theta) = \exp[\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}] \ , \ x \in \mathbb{R}$$

where $\theta := (\mu,\sigma^2)$. A natural parametrization is obtained by taking $\pi := (\mu/\sigma^2, -1/(2\sigma^2))$. Similarly, certain classes of Poisson, gamma and beta distributions constitute regular exponential families. □

In Theorem 1.9 admissibility of Bayes rules was considered and continuity of the risk function proved to be a major condition. The next theorem describes in detail when this condition is satisfied in the case of a one-parametric exponential family.

<u>Theorem 1.20</u>. Let Θ be an open (possibly infinite) interval in $\mathbb{R}$ and suppose that the following three conditions hold.

(i) Functions B_1 and $B_2 : \Theta \times \Theta \to \mathbb{R}$ exist, bounded on compact sets, such that

$$|L(\theta_2,a)| \leq B_1(\theta_1,\theta_2)|L(\theta_1,a)| + B_2(\theta_1,\theta_2)$$

holds for all θ_1 and $\theta_2 \in \Theta$;

(ii) $L(\theta,a)$ is continuous in θ for all $a \in A$;

(iii) $\{P_\theta,\theta \in \Theta\}$ is a one-parametric exponential family with density (1.13), where π is continuous and nondecreasing.

Then $R(\theta,d)$ is continuous in θ for every $d \in D$. $\square$

The basic feature of the proof is the straightforward inequality

$$|R(\theta_0,d)-R(\theta,d)| \leq \left|\int [L(\theta_0,d(x))-L(\theta,d(x))]f(x|\theta_0)\mu\{dx\}\right|$$

$$+ \int \left|L(\theta,d(x))[f(x|\theta_0)-f(x|\theta)]\right|\mu\{dx\}$$

For $|\theta-\theta_0| < \delta$ both terms on the right-hand side are bounded by integrable functions. Application of Lebesgue's bounded convergence theorem shows that the left-hand side tends to zero as θ tends to θ_0. See FERGUSON 1967, Theorem 3.7.2.

If will often be assumed in the sequel that the exponential family has been put into the canonical form; consequently, (1.13) will often be used with vector π of parameters replaced by θ.

2. TRUNCATED ESTIMATION PROBLEMS

2.1. Introduction and summary

In the chapters to follow the estimation theory, treated briefly in the previous chapter, will be applied to problems where certain parameter values can be excluded beforehand, reducing the class of possible probability distributions $\{P_\theta : \theta \in \Theta\}$. Often, such a reduced parameter space leads to a reduction of the action space as well. Estimation problems with a reduced action space will be discussed now.

Definition 2.1. Consider estimation problem (Θ,L,X) of Definition 1.4, where X has a probability distribution belonging to class $\{P_\theta : \theta \in \Theta\}$. Let Θ_0 be a strict subset of Θ. Then problem (Θ_0,L,X), where now L is restricted to Θ_0 and X has a distribution belonging to class $\{P_\theta : \theta \in \Theta_0\}$, is said to have a truncated parameter space (with respect to Θ). If $C\{h(\Theta_0)\}$ is a strict subset of $C\{h(\Theta)\}$, problem (Θ_0,L,X) is called a truncated estimation problem. □

The definition implies that for a truncated estimation problem a $\theta \in \Theta$ exists, such that $h(\theta) \notin C\{h(\Theta_0)\}$; hence, $h(\Theta_0)$ is a strict subset of $h(\Theta)$. Note that a truncated parameter space not necessarily leads to a truncated estimation problem; whether this occurs depends on the nature of the function h. Since a reduction of the parameter space is only of interest to the statistician in so far as the action space is reduced as well, the definition was chosen accordingly. In the sequel many applications will concern an exponential family; if it has the canonical form, any truncated parameter space necessarily is a subset of the natural parameter space.

The most important argument for the study of truncated parameter spaces is their rather frequent occurrence in practice. If one concentrates on exponential families, then certain parameter values in the natural parameter space can fairly often be excluded beforehand. Sometimes this exclusion follows from logical arguments as in the introductory example on randomized response. In other situations, theoretical and/or practical considerations about the nature of the specific problem at hand can lead to the exclusion of certain subsets of parameter values. If the statistician decides upon such a reduction of the parameter space for subjective reasons, his approach is in fact Bayesian: all prior distributions he is willing to take into consideration have their support inside the truncated parameter space.

Truncation of the parameter space reflects the statistician's prior knowledge about the parameter. The smaller the parameter space, the lesser the statistician's ignorance and the more accurate his estimates. The great practical importance of truncated parameter spaces therefore is that they allow estimators with a lower risk than could be achieved without truncation.

Truncated estimation problems give rise to specific mathematical and statistical problems. For example, the frequently used additional selection criterion of unbiasedness often cannot be applied, simply because no unbiased estimator exists; see Section 2.2 for precise conditions. Even more characteristic is the fact that boundary rules are inadmissible. (As stated in the Introduction, 'boundary rule' is the colloquial and rather vague name for an estimator taking values near the boundary of the action space.) Section 2.3 illustrates this feature for some special cases. The final Section 2.4 discusses some properties of minimax and Bayes rules. It is shown by means of examples that the risk of optimal rules indeed decreases, if the parameter space is truncated. However, minimax rules are more complicated than in the classical case and harder to derive. Among the examples are the results of CASELLA & STRAWDERMAN 1981 and BICKEL 1981.

Finally, a word of warning is appropriate on terminological issues. The term 'truncated' in the sense of Definition 2.1 goes back at least as far as KATZ 1961. However, the word 'restricted' is used as well, for example in the two references of the preceding paragraph. Furthermore, 'truncated' is in use as synonym for 'censored' to indicate observationary processes where no observation can exceed a given value, although the phenomenon itself can; this fully different notion will not occur here.

2.2. Absence of unbiased estimators

Unbiasedness undoubtedly is the most frequently used additional selection criterion for estimators. It should only be used, however, in combination with the criterion of admissibility, since in itself unbiasedness is no guarantee for a useful estimator. Some examples of absurd unbiased estimators can be found in FERGUSON 1967, p. 135 ff.

The number of unbiased estimators for a given problem typically is rather small; if a complete sufficient statistic exists, there is at most one nonrandomized unbiased estimator. If, in addition, the estimation problem is truncated, often no unbiased estimators exist. The following two theorems show the details.

Theorem 2.2. Consider estimation problem (Θ_0,L,X) that is truncated with respect to Θ. Let a statistic T exist which is sufficient for θ, $\theta \in \Theta$ and complete sufficient for θ, $\theta \in \Theta_0$, while the support of the distribution of T is the same for all $\theta \in \Theta$. Assume that a nonrandomized unbiased estimator exists for $h(\theta)$, $\theta \in \Theta$. Then no nonrandomized unbiased estimator exists for $h(\theta)$, $\theta \in \Theta_0$.

Proof. Denote the nonrandomized unbiased estimator for $h(\theta)$, $\theta \in \Theta$ by d and assume that a similar estimator d_0 exists for $h(\theta)$, $\theta \in \Theta_0$; since T is sufficient for θ, $\theta \in \Theta$, both estimators may be assumed to be based on T. (Replace any $d \in D$, not based on T, by the dominating rule $E\{d(X)\,|T\}$.) The completeness of T implies from

$$E_\theta[d(T) - d_0(T)] = 0 \text{ for all } \theta \in \Theta_0$$

that $d(T) = d_0(T)$ a.e. with respect to P_θ, $\theta \in \Theta_0$. Take $\theta' \in \Theta$ such that $h(\theta') \notin H_0 := C\{h(\Theta_0)\}$. If $P_{\theta'}\{d(T) \in H_0\} = 1$, the convexity of H_0 implies that $E_{\theta'}d(T) \in H_0$, which however contradicts the unbiasedness of d; therefore

$$P_{\theta'}\{d(T) \in H_0\} < 1$$

Now the support of the distribution of T does not depend on θ, hence

$$P_\theta\{d(T) \notin H_0\} > 0 \text{ for all } \theta \in \Theta$$

But this means that $d = d_0$ is not a function into H_0, hence no estimator at all for $h(\theta)$, $\theta \in \Theta_0$. □

Theorem 2.3. Consider estimation problem (Θ, L, X) with $\{P_\theta : \theta \in \Theta\}$ an exponential family. Let (Θ_0, L, X) be a truncated estimation problem with $\{P_\theta : \theta \in \Theta_0\}$ a regular exponential family. Assume that a nonrandomized unbiased estimator exists for $h(\theta)$, $\theta \in \Theta$. Then no nonrandomized unbiased estimator exists for $h(\theta)$, $\theta \in \Theta_0$.

Proof. Regularity of $\{P_\theta : \theta \in \Theta_0\}$ ensures that Θ_0 is open. So, a statistic T exists which is sufficient for θ, $\theta \in \Theta$ and complete sufficient for θ, $\theta \in \Theta_0$; moreover, the support of T is the same for all $\theta \in \Theta$. Now, application of Theorem 2.2 completes the proof. □

An important application of Theorem 2.3 concerns the situation where Θ is the natural parameter space of a regular k-parametric exponential family. If an unbiased estimator exists for $h(\theta)$, $\theta \in \Theta$, then no unbiased estimator exists for $h(\theta)$, $\theta \in \Theta_0$, whenever Θ_0 contains an open set in $\mathbb{R}_k$.

The proof of Theorems 2.2 and 2.3 is based on the fact that any unbiased estimator d for $h(\theta)$, $\theta \in \Theta$ cannot be an estimator for $h(\theta)$, $\theta \in \Theta_0$ as well. An obvious way to turn d into an estimator for $h(\theta)$, $\theta \in \Theta_0$ would be to project (perpendicularly) all estimates $d(x)$ on $C\{h(\Theta_0)\}$. The resulting estimator will be biased, however; in fact, it will not even be admissible in general, as the next section will show.

2.3. Inadmissibility of boundary rules

In this section it will be shown that estimators taking values upon or close to the boundary of the action space of a truncated convex estimation problem are often inadmissible. This feature is specific for truncated parameter spaces; for instance, it does not occur in natural parameter spaces. The discussion starts with two simple examples, where the action space is an interval and quadratic loss is used.

Example 2.4. Let the truncated quadratic estimation problem (Θ,L,X) be defined by $X \in B(1,\theta)$, $L(\theta,a) = (\theta-a)^2$ and $\Theta = (Q,P) \subset (0,1)$, where Q and P ($> Q$) are given constants. The maximum likelihood estimator (MLE) d_1 for this problem is given by $d_1(0) = Q$ and $d_1(1) = P$; its risk function equals

$$R(\theta,d_1) = \theta(\theta-P)^2 + (1-\theta)(\theta-Q)^2 \tag{2.1}$$

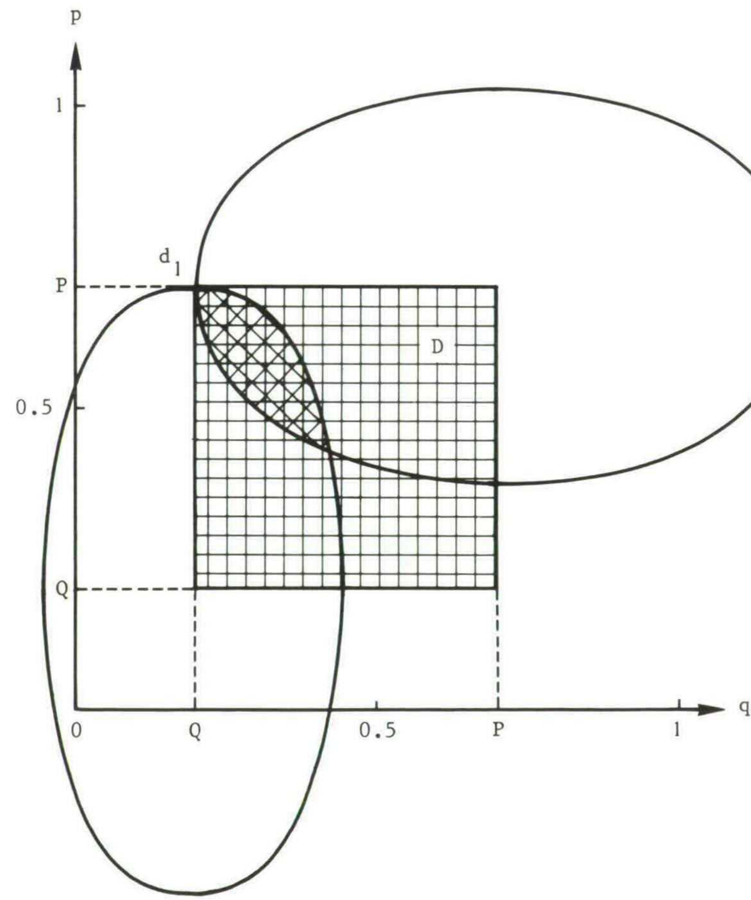


Figure 2.5. Set of estimators dominating d_1 (doubly shaded area) in Example 2.4

Although d_1 looks rather plausible, it is an inadmissible estimator, except in the cases $Q = 0$ or $P = 1$ (one-sided truncation). To see this, define d_{qp} by $d_{qp}(0) = q$ and $d_{qp}(1) = p$ with $Q \leqslant q \leqslant p \leqslant P$. The risk of d_{qp} is analogous to (2.1) and it is easy to check that $F(\theta) := R(\theta,d_1) - R(\theta,d_{qp})$ is a parabola with a maximum. Now, d_1 is dominated by d_{qp}, if and only if $F(\theta)$ is nonnegative on the whole interval $[Q,P]$; this is the case if and only if $F(Q)$ and $F(P)$ both are nonnegative, or equivalently, if the following two inequalities are satisfied:

$$(2.2) \qquad \begin{cases} Q(p-Q)^2 + (1-Q)(q-Q)^2 \leqslant Q(P-Q)^2 \\ \\ P(p-P)^2 + (1-P)(q-P)^2 \leqslant (1-P)(P-Q)^2 \end{cases}$$

In the (q,p)-plane, the inequalities are represented by two ellipses with midpoints (Q,Q) and (P,P), respectively; their intersection represents the set of all (q,p)-values satisfying (2.2), i.e. all estimators d_{qp} that strictly dominate d_1. It is the shaded area in Figure 2.5. So, for any pair (Q,P) with $0 < Q < P < 1$, d_1 is inadmissible. Note that for $Q = 0$, $Q = P$ or $P = 1$ an ellipse collapses. □

This simple example illustrates that estimators taking values exactly on the boundary of the action space may be inadmissible. This will often occur to the MLE in truncated estimation problems. For example, let the class of probability distributions be a regular exponential family in the canonical form. Then the MLE for θ, $\theta \in \Theta$ either is on the boundary of Θ or satisfies - by (1.14) - the first order condition

$$(2.3) \qquad \nabla c(\theta) = t(x)$$

The likelihood indeed attains a (local) maximum for a solution of (2.3), as its Hessian equals $- Hc(\theta)$, which is strictly negative definite: compare (1.15). Since (2.3) does not depend on the choice of Θ, truncated parameter spaces exist, which do not contain the solution(s) of (2.3); in that case the MLE necessarily is on the boundary.

Furthermore, almost all numerical procedures for the implementation of estimation methods replace estimates that would fall outside the closed convex parameter space by the nearest point within that space. In the case of inequality constraint regression, for example, this is the common procedure; compare LIEW 1976, DAVIS 1978, CHANG 1981 and DYKSTRA 1983, among many others. Estimating the probabilities of a Markov process offers another example. Since the probabilities must be nonnegative and sum to 1, this problem can be viewed as a truncated estimation problem. All numerical methods supply the value 0, whenever a negative estimate should occur; e.g. see LEE et al. 1977. In all these cases, the resulting estimators generally will be inadmissible.

Example 2.4 illustrated that rules taking values precisely on the boundary of the action space, can be inadmissible. From continuity arguments it seems plausible that in fact even rules with values close to the boundary often will be inadmissible. This will be shown next for the estimation problem of Example 2.4.

Example 2.6. Consider again the truncated estimation problem of Example 2.4, in particular the rules d_{qp}. For any (q,p) with $Q \leqslant q \leqslant p \leqslant P$ estimators d_{qpr} for $\theta \in \Theta$ are defined by $d_{qpr}(0) = q + r$, $d_{qpr}(1) = p-r$; further, $D_{qpr} := \{d_{qpr} : 0 \leqslant r \leqslant (p-q)/2\}$. Now, d_{qp} will be compared with the other estimators in D_{qpr}. It is clear that

$$
\begin{aligned}
F_r(\theta) &:= R(\theta,d_{qp}) - R(\theta,d_{qpr}) \\
&= \theta(\theta-p)^2 + (1-\theta)(\theta-q)^2 - \theta(\theta-p+r)^2 - (1-\theta)(\theta-q-r)^2 \\
&= -\theta[r^2 + 2r(\theta-p)] - (1-\theta)[r^2 - 2r(\theta-q)] \\
&= r[-4\theta^2 + 2\theta(1+p+q) - r - 2q]
\end{aligned}
$$

represents a parabola having a maximum. Hence, $F_r(\theta)$ is nonnegative on the whole interval $[Q,P]$, if and only if $F_r(Q)$ and $F_r(P)$ are both nonnegative, or equivalently,

$$
(2.4) \qquad \begin{cases} r/2 \leqslant -2Q^2 + Q(1+p+q) - q \\ r/2 \leqslant -2P^2 + P(1+p+q) - q \end{cases}
$$

An $r > 0$ satisfying (2.4) only exists if both right-hand sides are positive, which is equivalent to the condition that p and q satisfy (2.5):

$$(2.5) \qquad p > \max[q(1-P)/P + 2P - 1,\ q(1-Q)/Q + 2Q - 1]$$

Summing up: under condition (2.5), an $r > 0$ exists such that $F_r(\theta)$ is positive uniformly on $[Q,P]$; then d_{qp} is dominated strictly by the corresponding d_{qpr} and therefore inadmissible. The doubly shaded area in Figure 2.7 shows the set of (q,p)-values satisfying (2.5); so, this area represents the inadmissible rules d_{qp}.

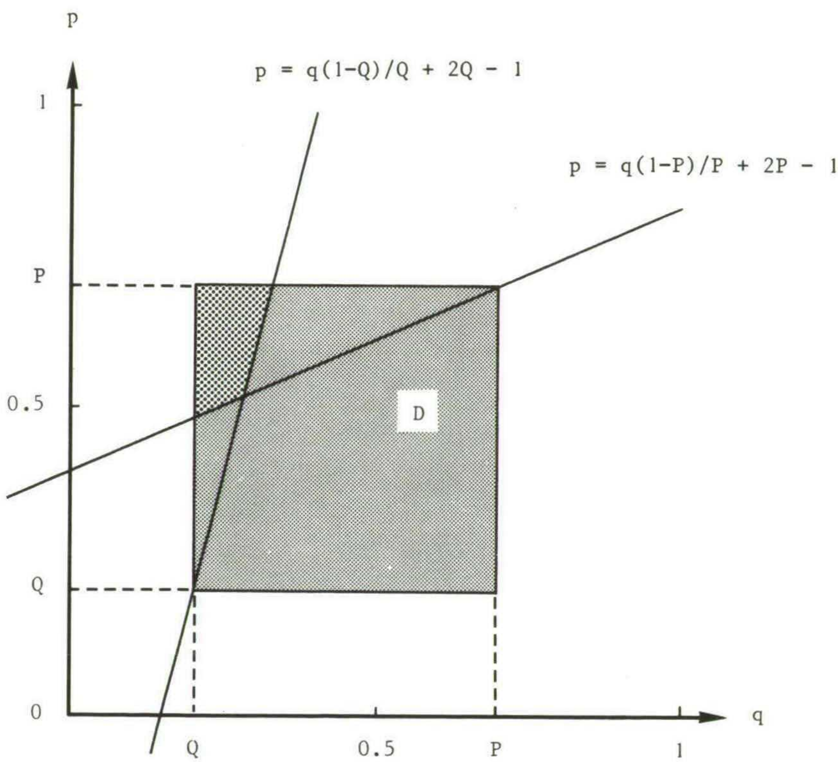


Figure 2.7. Set of inadmissible estimators d_{pq} (doubly shaded area) in Example 2.6

Point of intersection of the two lines shown in Figure 2.7 is (2PQ, 2(P+Q-PQ)-1); it lies inside the shaded square for $0 < Q < \frac{1}{2} < P < 1$. Hence, in this situation there always exist (q,p)-values for which d_{qp} is inadmissible. If, however, Q and P are on the same side of $\frac{1}{2}$, none of the estimators d_{qp} is strictly dominated by any rule in D_{qpr}. As final numerical application, take $\Theta = (0.2,0.7)$; then estimates 0.25 and 0.6 for $x = 0$ and $x = 1$, respectively, combine to an inadmissible estimator. Yet, this estimator seems perfectly plausible at first sight. □

From these two examples it can be concluded that not only rules taking values precisely on the boundary of the action space can be inadmissible, but even - at first sight - plausible rules with values in the interior. An obvious but important exception is constituted by the class of so-called constant rules: rules that do not depend on the observations. Any rule d taking a constant value $d(X) = c \in h(\Theta)$ is admissible, because its risk is zero if c happens to coincide with the true value of $h(\theta)$. Constant rules are extremely unsatisfactory, on the other hand, since the observations become irrelevant; in fact, the statistical problem is reduced to a deterministic one. The class of constant rules therefore constitutes an awkward category, consisting of unsatisfactory but admissible rules. This paradox can be viewed as a weak spot in the concept of admissibility. It could be removed by excluding constant rules, e.g. by requiring that any admissible estimator satisfies the information inequality (if applicable). However, rather than modifying the standard concept of admissibility, starting in Chapter 3 constant rules will be excluded by imposing invariance as additional optimality criterion.

2.4. Minimax and Bayes estimators

The minimax principle for nonrandomized estimators was defined in Section 1.4; according to this criterion, estimators are ordered with respect to their maximum risk. This induces a complete ordering in D and therefore leads to a 'best' estimator. It was shown previously, that minimax estimators exist under fairly general conditions (Theorem 1.10).

Generally, minimax rules are hard to find; their calculation often constitutes a tough (numerical) problem. Sometimes it can be tackled successfully by first deriving a least favorable prior distribution as well as the corresponding Bayes rule; its risk exactly equals the minimax risk, provided the estimation problem has a value (Theorem 1.11). So, there is a natural correspondence between minimax rules on the one hand and Bayes rules with respect to least favorable priors on the other.

The derivation of Bayes estimators is relatively easy. Denote by Z a random variable on Θ with distribution $\tau \in \Theta^*$. Provided that an estimator exists with finite risk and provided that for almost all $x \in X$ a value $d_\tau(x)$ exists minimizing $E_\tau\{L(Z,d(X))|X = x\}$, the resulting estimator d_τ is Bayes with respect to the prior distribution $\tau \in \Theta^*$. For the quadratic loss function (1.3), the Bayes estimator then becomes the expectation of the posterior distribution of the estimand:

$$(2.6) \qquad d_\tau(x) = E_\tau\{h(Z)\big| X = x\}$$

Generally, τ will be dominated by some σ-finite measure ν on Θ; without fear of confusion, its density can be denoted by τ as well. If class $\{P_\theta : \theta \in \Theta\}$ is dominated too - with respect to μ and with densities $f(x|\theta)$ -, the joint density $f(x,\theta)$ equals $f(x|\theta)\tau(\theta)$. In this case (2.6) may be rewritten as

$$(2.7) \qquad d_\tau(x) = \int_\Theta h(\theta)f(x,\theta)\nu\{d\theta\}\Big/\int_\Theta f(x,\theta)\nu\{d\theta\}$$

which will be used later on. It will be assumed that density $f(x,\theta)$ is measurable in the pair (x,θ); then estimator d_τ is Borel measurable.

Class Θ^* of all prior distributions is too large to enable a complete analysis: restricting attention to a suitable smaller class is desirable. A good choice are the so-called conjugated families, introduced by RAIFFA & SCHLAIFER 1961; compare also BARNDORFF-NIELSEN 1978 and DIACONIS & YLVISAKER 1979.

Definition 2.8. Family $\{\tau_\alpha : \alpha \in A\}$ of probability distributions of θ is called conjugated to class $\{P_\theta : \theta \in \Theta\}$ of distributions of X, if the family is closed under conditioning with respect to X. □

Obviously, the great advantage of a prior distribution $\tau_\alpha(\theta)$ from a conjugated family is that the posterior distribution $\tau_\alpha(\theta|x)$ belongs to the same family for all $x \in \mathcal{X}$. For any exponential family, a conjugated family is an exponential family itself; for the exponential family with densities (1.13), the priors of a conjugated family may be notated as

$$(2.8) \qquad \tau_\alpha(\theta) = \exp[\sum_{i=1}^{k} \alpha_i \pi_i(\theta) - \alpha_{k+1} c(\theta) + v(\alpha)] \ , \ \theta \in \Theta$$

with $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_k, \alpha_{k+1}) \in A \subset \mathbb{R}_{k+1}$. For τ to be a proper probability distribution the normalizing function $v : A \to \mathbb{R}$ has to satisfy the condition $\int \tau_\alpha(\theta)\nu\{d\theta\} = 1$. Of course, a conjugated family is not unique.

Of special interest is the so-called noninformative (or Jeffreys) prior, reflecting absence of prior knowledge about θ. In general, it is the uniform measure on Θ, corresponding to $\alpha = 0$ in (2.8). Therefore, in the sequel A will be assumed to contain the 0-vector. Note that the corresponding τ_0 may be a probability distribution only if Θ is finite.

Let Θ_0 be a truncated parameter space with respect to Θ. Then conjugated families to $\{P_\theta : \theta \in \Theta\}$ and $\{P_\theta : \theta \in \Theta_0\}$ respectively will be closely related. Since in particular the support of τ_α differs between these cases, the main difference will be a normalizing function of α only. For exponential families this concerns the function $v(\alpha)$ in (2.8). Example 2.10 presents an illustration; it uses the truncated beta distributions of Definition 2.9.

Definition 2.9. A distribution defined for $\frac{1}{2} < P < 1$ by the density

$$(2.9) \qquad \tau^*_{(\alpha,\beta)}(\theta) = C^{-1}(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1} I_{[1-P,P]}(\theta) \ , \ \alpha \in \mathbb{R} \ , \ \beta \in \mathbb{R}$$

where

$$(2.10) \qquad C(\alpha,\beta) := \int_{1-P}^{P} \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$

is called a (symmetrically) truncated beta distribution. □

The distribution with density (2.9) will be denoted by $B(\alpha,\beta;P)$. Note that they are defined for negative values of α and β as well, contrary to the usual nontruncated beta distributions.

Example 2.10. First, consider the exponential family of binomial distributions $\{B(n,\theta) : 0 < \theta < 1\}$. The class of beta distributions $\{Be(\alpha,\beta) : \alpha > 0,\ \beta > 0\}$ is conjugated to this family. Indeed,

$$\tau_{(\alpha,\beta)}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\ ,\ 0 < \theta < 1$$

implies

$$\tau_{(\alpha,\beta)}(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\ ,\ 0 < \theta < 1$$

where $\propto$ denotes equality up to a factor independent of θ; so a posteriori a beta distribution arises again. Note that the family $\{Be(\alpha,\beta) : \alpha \in \mathbb{N},\ \beta \in \mathbb{N}\}$ is conjugated to $\{B(n,\theta) : 0 < \theta < 1\}$ as well, showing that more than one conjugated family may exist with respect to a given exponential family.

Now consider the binomial distributions with (symmetrically) truncated parameter space: $\{B(n,\theta) : 1-P < \theta < P\}$, where $\frac{1}{2} < P < 1$. A similar reasoning as above holds for $\tau^*_{(\alpha,\beta)}$ in (2.9), showing that the family of truncated beta distributions is conjugated to $\{B(n,\theta) : 1-P < \theta < P\}$. □

For easy reference the last part of this example is stated as a seperate lemma.

Lemma 2.11. For any P with $\frac{1}{2} < P < 1$, family $\{B(\alpha,\beta;P) : \alpha \in \mathbb{R},\ \beta \in \mathbb{R}\}$ is conjugated to family $\{B(n,\theta) : 1-P < \theta < P\}$. □

As stated before, minimax rules are hard to derive in general. The problem gets even harder in truncated estimation problems. The next two examples illustrate this statement.

Example 2.12. Consider truncated quadratic estimation problem (Θ,L,X) with $X \in B(1,\theta)$, $L(\theta,a) = (\theta-a)^2$ and $\Theta = (1-P,P)$, $\frac{1}{2} < P < 1$. This is a special case of the problem considered in Example 2.4, now with the symmetric truncation $Q = 1-P$. As in Example 1.14, the problem is invariant under $\{e,g\}$, where $g(x) = 1-x$. A minimax estimator d_m can be found within the class of invariant estimators (Theorem 1.12); hence it should satisfy $d_m(1-x) = 1-d_m(x)$. Define invariant estimators d_p by $d_p(1) = p$, for $\frac{1}{2} < p \leqslant P$; the risk equals

$$(2.11) \qquad \begin{aligned} R(\theta,d_p) &= \theta(\theta-p)^2 + (1-\theta)(\theta-(1-p))^2 \\ &= \theta(1-\theta)(4p-3) + (1-p)^2 \end{aligned}$$

For $p < \frac{3}{4}$ the right-hand side represents a parabola having a minimum; hence

$$(2.12) \qquad \max_\theta R(\theta,d_p) = R(P,d_p) = (1-\phi)(p-\tfrac{3}{4}) + (1-p)^2$$

where $\phi := (2P-1)^2$. The minimum of (2.12) equals $\phi(1-\phi)/4$ and is attained for $p = (1+\phi)/2$. It is easy to check that for $p \geqslant \frac{3}{4}$ the minimax value of (2.11) equals 1/16, attained for $p = \frac{3}{4}$. Comparison of these two cases shows that the minimax estimator is given by

$$(2.13) \qquad d_m(1) = \begin{cases} (1+\phi)/2 & \text{for } \phi \leqslant \frac{1}{2} \\ 3/4 & \text{for } \phi \geqslant \frac{1}{2} \end{cases}$$

The maximum risk connected with this rule is

$$\max_\theta R(\theta,d_m) = \begin{cases} \phi(1-\phi)/4 & \text{for } \phi \leqslant \frac{1}{2} \\ 1/16 & \text{for } \phi \geqslant \frac{1}{2} \end{cases}$$

Note that minimax rule (1.12) for the nontruncated case is included in (2.13). Figure 2.13 shows the risk functions of the minimax rule in the cases $P = \frac{3}{4}$ (where $d_m = d_p$ with $p = 5/8$), $P = 0.83$ (with $d_m(1) = 0.7178$) and $P = 1$.

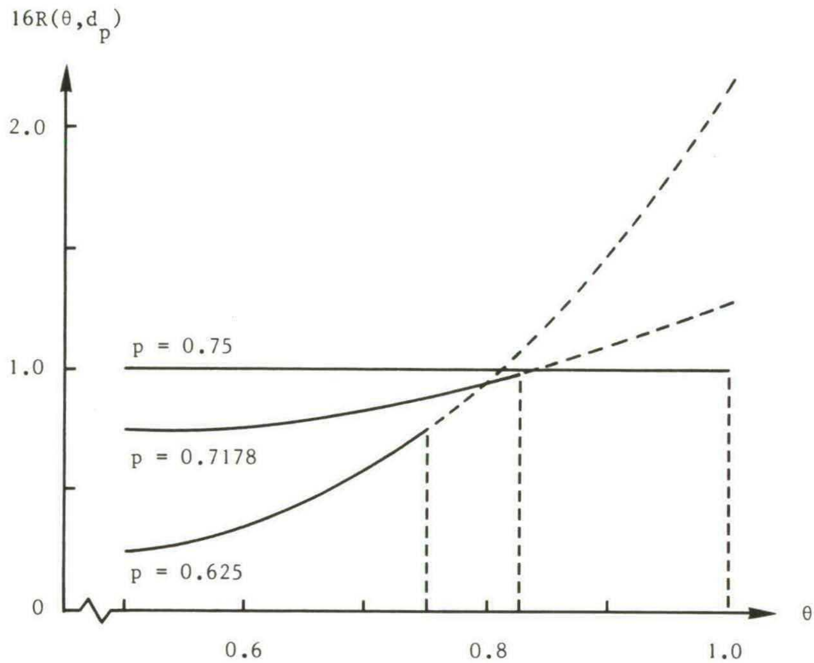


Figure 2.13. Risk functions of minimax estimators for binomial parameter for various truncations

A least favorable prior τ_* can be found among the invariant distributions on Θ (Theorem 1.13), in this case the symmetric distributions around $\frac{1}{2}$. From the general expression $\tau(\theta|1) \propto \theta\tau(\theta)$, holding for any $\tau \in \Theta^*$, follows

$$\tau(\theta|1) = 2\theta\tau(\theta), \quad \theta \in \Theta$$

for any invariant $\tau \in \Theta^*$. Now (2.6) implies that the invariant Bayes rule d_τ must satisfy $d_\tau(1) = 2E_\tau(Z^2)$, where Z has distribution τ on Θ. The corresponding Bayes risk is obtained by taking the expectation of (2.11) with respect to τ:

$$r(\tau,d_\tau) = [\tfrac{1}{2} - E_\tau(Z^2)][8E_\tau(Z^2) - 3] + [1 - 2E_\tau(Z^2)]^2$$

$$= \mathrm{Var}_\tau(Z)[1 - 4\ \mathrm{Var}_\tau(Z)]$$

Maximizing this expression leads to the least favorable invariant prior:

$$\tau_* = \begin{cases} \tau_{-\infty} & \text{for } \phi \leq \tfrac{1}{2} \\ \text{any symmetrical } \tau \text{ with } \mathrm{Var}_\tau(Z) = 1/8 & \text{for } \phi \geq \tfrac{1}{2} \end{cases}$$

where $\tau_{-\infty}$ is the two-point distribution, giving equal probability to the endpoints of Θ : $P\{Z = 1-P\} = P\{Z = p\} = \tfrac{1}{2}$. An example of a least favorable distribution for $\phi > \tfrac{1}{2}$ is the three-point distribution defined by $P\{Z = 1 - P\} = P\{Z = P\} = 1/(4\phi)$ and $P\{Z = \tfrac{1}{2}\} = 1 - 1/(2\phi)$. □

<u>Example 2.14</u>. Minimax estimators for the mean of a symmetrically truncated normal distribution were considered by CASELLA & STRAWDERMAN 1981. For problem (Θ,L,X) with $X \in N(\theta,1)$, $L(\theta,a) = (\theta-a)^2$ and $\Theta = [-m,m]$ they proved the following.

(i) For $0 < m \leq 1.057$ the minimax rule d_m is given by

$$d_m(x) = m \tanh(mx) \tag{2.14}$$

Furthermore, a corresponding least favorable prior is the two-point distribution determined by $P\{\theta = \pm m\} = \tfrac{1}{2}$.

(ii) For $1.4 \leq m \leq 1.6$ the minimax rule is

$$d_m(x) = \frac{(1-\alpha)m\ \tanh(mx)}{1-\alpha+\alpha\ \exp(m^2/2)/\cosh(mx)}$$

where α is uniquely determined by the additional condition $R(0,d_m) = R(m,d_m)$. A least favorable prior now is the three-point distribution defined by $P\{\theta = \pm m\} = (1-\alpha)/2$ and $P\{\theta = 0\} = \alpha$. From numerical evidence the authors conjectured that this result holds in fact on a wider interval of m-values. Figure 2.15 presents the risk functions of the minimax rules for $m = 1$ and $m = 1.5$. For comparison the risk functions of the MLE are drawn as well; of course, the MLE is the rule defined by

$$d(x) = \begin{cases} -m \text{ for } x < -m \\ x \text{ for } |x| \leq m \\ m \text{ for } x > m \end{cases}$$

It is easy to see that for $m \to \infty$ the observation X itself is minimax with constant risk 1.

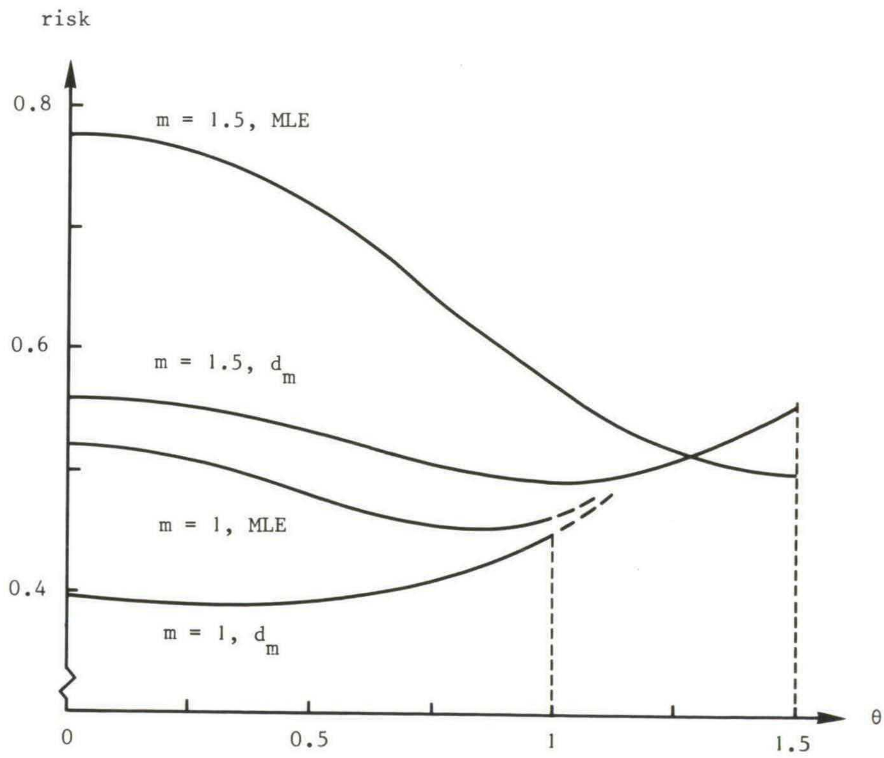


Figure 2.15. Risk functions of minimax estimators and MLE's for truncated normal distribution

BICKEL 1981 considered the same problem, for large values of m. He proved the density

$$(2.15) \qquad \tau_*(\theta) = \frac{1}{m} \cos^2\left(\frac{\pi}{2m}\theta\right), \quad -m \leq \theta \leq m$$

to be a least favorable prior, approximately. The corresponding maximum risk equals $1 - \pi^2/m^2 + o(m^{-2})$.

ZEYTINOGLU & MINTZ 1984 considered this problem in the testing situation, i.e. with a zero-one loss function; since this loss function is not convex, their problem is not regular in the sense of Definition 1.5. □

Note the striking similarity between these two examples as regards the least favorable prior. For small parameter spaces, it is the two-point distribution, having point masses $\frac{1}{2}$ at the endpoints; for wider intervals, an additional point mass at the midpoint appears. More generally, in truncated decision problems a least favorable prior distribution exists, which is concentrated in a limited number of points. In Chapter 6 this feature will be discussed in more detail for the binomial distribution.

3. INVARIANT ESTIMATION PROBLEMS

3.1. Introduction and summary

In Section 2.3. some examples were presented illustrating the fact that quite plausibly looking boundary rules can be inadmissible. It was noted, that these examples are difficult to generalize to a theory stating exactly what kind of boundary rules are inadmissible under what conditions. Of the complications that arise in such an attempted generalization, the existence of constant rules was the most annoying: they are always admissible, even if their value is on the boundary of the action space. This complication will be eliminated here by restricting attention to invariant estimators, defined in Section 1.4. Hence the general framework of an estimation problem, as presented in Definition 1.4 will be curtailed by only considering invariant estimation problems. In fact, some additional assumptions will be made, to be discussed now. Note that the exposition about to follow is not restricted to truncated estimation problems, but applies to the classical case as well. However, the more interesting applications will be seen to exist in the field of truncated estimation theory.

An estimation problem (Θ,L,X) will be considered here, which is invariant under a finite group G of functions $g : \mathcal{X} \to \mathcal{X}$, while the class $\{P_\theta : \theta \in \Theta\}$ of probability distributions of X is dominated by some measure μ with density $f(x|\theta)$. Three further assumptions will be made about the invariance.

First of all, it will be supposed that all $g \in G$ are <u>measure preserving</u>, which means that

$$(3.1) \qquad \mu\{g^{-1}(B)\} = \mu\{B\}$$

holds for all Borel sets $B \subset \mathcal{X}$. (Note that g^{-1} exists and belongs to G, since G is a group.) For the counting measure this property is trivially satisfied. If μ is the Lebesgue measure, simple functions g, like translations, rotations and reflexions, have this property. See KINGMAN & TAYLOR 1966, Section 7.5 or ZAANEN 1967, Section 5.8. That (3.1) is not true for any group G is shown by means of a simple counterexample. For any space S, the function $e : S \to S$ will denote the identity.

Example 3.1. Let X be the half open unit interval $[0,1)$; μ is the Lebesgue measure. Define $g : X \to X$ by

$$g(x) := \begin{cases} 2x + 1/3 & \text{for } x \in [0,1/3) \\ (x - 1/3)/2 & \text{for } x \in [1/3,1) \end{cases}$$

Then $gg = e$, so that $\{e,g\}$ is a group. But g is not measure preserving: g maps $[0,1/3)$ onto $[1/3,1)$, thereby doubling its (Lebesgue) measure. □

The last assumptions concern the group $\tilde{G}$. Recall that G induces groups $\bar{G}$ and $\tilde{G}$ of functions $\bar{g} : \Theta \to \Theta$ and $\tilde{g} : A \to A$, respectively. It will be assumed that all functions $\tilde{g} \in \tilde{G}$ are linear, meaning here that

$$\tilde{g}(\alpha x_1 + \beta x_2) = \alpha\tilde{g}(x_1) + \beta\tilde{g}(x_2)$$

for all α and $\beta \in \mathbb{R}$ and all x_1 and $x_2 \in A$ such that $\alpha x_1 + \beta x_2 \in A$. Finally, it will be assumed that $\tilde{G}$ is commutative. These assumptions are summarized in the following definition.

Definition 3.2. Let estimation problem (Θ,L,X) be invariant under a finite group G with induced group $\tilde{G}$ of functions $\tilde{g} : A \to A$; let $\{P_\theta : \theta \in \Theta\}$ be dominated by μ. If all $g \in G$ are measure preserving, $\tilde{G}$ is commutative and all $\tilde{g} \in \tilde{G}$ are linear, the problem is called linearly invariant or linvariant for short. Invariant estimators for this problem will be called linvariant as well. □

Until the final Section 3.6, the quadratic loss function (1.3) will be used almost exclusively, so that linvariant quadratic estimation problems will be the object of study here. Note that this framework is not extremely restrictive: most familiar invariant estimation problems belong to this class.

The main results obtained can be informally stated as follows. For any observation $x \in X$, the action space $A = C\{h(\Theta)\}$ can in fact be reduced to a subset A_x in such a way that only inadmissible rules are excluded. In other words, any rule taking values outside A_x (with positive probability for some $\theta \in \Theta$) is inadmissible. The important feature is of course that this statement holds regardless of the exact values of the rule inside A_x. By consequence, the action space may be restricted beforehand - though only conditionally on the observation $x \in X$ -, thereby limiting the class of potentially useful estimators. To be of any use, it is necessary that A_x is a strict subset of A, at least for some $x \in X$. It will be seen that this often occurs; in such a case the boundary area of A is excluded, generalizing the examples of Section 2.3.

To obtain these results a number of preparatory lemmas is needed, which are presented in Section 3.2. They disclose the deeper structure of the linvariant problems under consideration. Further, the subsets $A_x \subset A$ are defined and some of their properties highlighted; some results regarding perpendicular projections onto convex spaces are presented. The main Theorem 3.9 is stated and proved in Section 3.3. This result is considered the most interesting finding of this study; as stated before, it was published earlier (MOORS 1981b or its preliminary version MOORS 1981a). Further, the notion of an inadmissible (stochastic) set is introduced; roughly speaking, a set V is inadmissible, if any rule taking values in V with positive probability is inadmissible. By means of this concept, Theorem 3.9 can be reformulated and clarified.

Section 3.4 and 3.5 give various applications. Section 3.4 concentrates on classical, nontruncated decision problems, where the results are not very spectacular. More interesting examples can be found for the truncated parameter spaces, treated in Section 3.5. The application to regression theory, where the parameters are known to satisfy given linear inequalities, is especially enlightening. Since inequality constraint regression is quite a topic currently (compare TOUTENBURG 1982), these results seem to be of importance. Note that numerical solutions usually project possible estimates outside the truncated parameter space onto its boundary; the present findings indicate, however, that the resulting estimators are inadmissible.

The final Section 3.6 discusses possible generalizations of these results. Other loss functions are considered and the notion of pseudo-invariance is introduced.

3.2. Preliminary lemmas

Before stating and proving the main theorem, six preparatory lemmas are presented, which reveal the deeper structure of linvariant estimation problems. The first two lemmas are direct consequences of the definition of a linvariant problem. Recall that a functions $\tilde{g} : A \to A$ is called isometric if $\tilde{g}$ preserves distances, so that $|\tilde{g}(a)| = |a|$ for all $a \in A$.

Lemma 3.3. Let problem (Θ,L,X) be linvariant under G, with L quadratic; let any $g \in G$ induce functions $\bar{g} : \Theta \to \Theta$ and $\tilde{g} : A \to A$. Then for any $g \in G$

(i) $h\bar{g} = \tilde{g}h$ on Θ;

(ii) $\tilde{g}$ is isometric.

Proof. Take any $g \in G$. From the definition of the loss function and the invariance property (1.5) the equalities

$$(3.2) \qquad |h(\theta) - a|^2 = L(\theta,a) = L(\bar{g}(\theta),\tilde{g}(a)) = |h\bar{g}(\theta) - \tilde{g}(a)|^2$$

follow for all $a \in A$ and all $\theta \in \Theta$. Choosing $a = h(\theta)$ implies

$$|h\bar{g}(\theta) - \tilde{g}h(\theta)|^2 = 0 \ , \ \theta \in \Theta$$

proving (i). Substitution of this result in (3.2) gives

$$(3.3) \qquad |h(\theta) - a|^2 = |\tilde{g}h(\theta) - \tilde{g}(a)|^2$$

Now, A contains an open set in $\mathbb{R}_m$ by Definition 1.4; therefore, $\tilde{g}$ can be uniquely extended to a linear function on the entire $\mathbb{R}_m$, which can be represented by a m×m matrix M. Since $\tilde{g}$ and M are identical on A, relation

$$|\tilde{g}h(\theta) - \tilde{g}(a)|^2 = |Mh(\theta) - Ma|^2 = |M(h(\theta)-a)|^2$$

holds for all $\theta \in \Theta$ and all $a \in A$. Combination with (3.3) gives $|u|^2 = |M(u)|^2$ for all $u \in U$, where $u = h(\theta) - a$. Since U contains an open set in $\mathbb{R}_m$, M is orthonormal, so $\tilde{g}$ must be isometric. □

Lemma 3.4. Let problem (Θ,L,X) be linvariant under G. Then for any $g \in G$ and any $\theta \in \Theta$

$$(3.4) \qquad f(x|\bar{g}(\theta)) = f(g^{-1}(x)|\,\theta)$$

holds a.e. with respect to μ.

Proof. For any Borel set $B \subset X$, (1.4) and (3.1) give

$$\int_B f(x|\,\bar{g}(\theta))\mu\{dx\} = P_{\bar{g}(\theta)}\{X \in B\} = P_\theta\{X \in g^{-1}(B)\}$$

$$= \int_{g^{-1}(B)} f(y|\theta)\mu\{dy\} = \int_{g^{-1}(B)} f(y|\theta)\mu g^{-1}\{dy\}$$

$$= \int_B f(g^{-1}(x)|\,\theta)\mu\{dx\}$$

where the last equality is obtained by the change of variable $x = g(y)$. □

Note that the linearity of the functions $\tilde{g}$ is not used in the proof.

Next, for all $x \in X$, subsets $A_x \subset A$ will be defined. In a slightly more elaborate notation the functions $g \in G$ are supplied with a suffix $i \in I$ so that $G = \{g_i : i \in I\}$; throughout the chapter, Σ will indicate summation over all $i \in I$. With the aid of ratios

$$\alpha(x|\,\bar{g}_j(\theta)) := f(x|\bar{g}_j(\theta))/\Sigma f(x|\,\bar{g}_i(\theta))$$

functions $h_x : \Theta \to A$ are defined for any fixed $x \in X$ by

$$(3.5) \qquad h_x(\theta) := \begin{cases} \Sigma\alpha(x|\,\bar{g}_i(\theta))\tilde{g}_i h(\theta) & \text{for } \Sigma f(x|\,\bar{g}_i(\theta)) > 0 \\ h(\theta) & \text{for } \Sigma f(x|\,\bar{g}_i(\theta)) = 0 \end{cases}$$

Now, A_x is defined as the convex closure of the range of h_x:

$$(3.6) \qquad A_x := C\{h_x(\Theta)\}$$

The next two lemmas state some properties of the entities defined above.

Lemma 3.5. The functions h_x defined by (3.5) satisfy

$$h_{g(x)}\bar{g} = \tilde{g}h_x$$

on Θ for any $g \in G$.

Proof. Straightforward application of (3.5) in combination with the linearity of $\tilde{g}$ gives

$$h_{g(x)}\bar{g}(\theta) = \begin{cases} \Sigma\alpha(g(x)|\bar{g}_i\bar{g}(\theta))\tilde{g}_i h\bar{g}(\theta) & \text{for } \Sigma\alpha(g(x)|\bar{g}_i\bar{g}(\theta)) > 0 \\ h\bar{g}(\theta) & \text{for } \Sigma\alpha(g(x)|\bar{g}_i\bar{g}(\theta)) = 0 \end{cases}$$

$$\tilde{g}h_x(\theta) = \begin{cases} \Sigma\alpha(x|\bar{g}_i(\theta))\tilde{g}\tilde{g}_i h(\theta) & \text{for } \Sigma\alpha(x|\bar{g}_i(\theta)) > 0 \\ \tilde{g}h(\theta) & \text{for } \Sigma\alpha(x|\bar{g}_i(\theta)) = 0 \end{cases}$$

The commutativity of $\tilde{G}$ implies that of $\bar{G}$, because of Lemma 3.3(i). In combination with Lemma 3.4 this easily shows the equality of the two expressions above. □

Lemma 3.6. The spaces A_x are subsets of A satisfying

$$(3.7) \qquad A_{g(x)} = \tilde{g}(A_x)$$

for any $g \in G$.

Proof. Lemma 3.5 and the fact that $\bar{g}$ is a surjection imply

$$h_{g(x)}(\Theta) = h_{g(x)}\bar{g}(\Theta) = \tilde{g}\{h_x(\Theta)\}$$

Now (3.7) results from the following two observations:

(i) any a in the convex hull of $h_x(\Theta)$ can be written as $\alpha a_1 + (1-\alpha)a_2$ with a_1 and $a_2 \in h_x(\Theta)$ and $0 \leq \alpha \leq 1$; then $\tilde{g}(a) = \alpha\tilde{g}(a_1) + (1-\alpha)\tilde{g}(a_2)$, where $\tilde{g}(a_1)$ and $\tilde{g}(a_2)$ both belong to $h_{g(x)}(\Theta)$, so that $\tilde{g}(a)$ is in the convex hull of $h_{g(x)}(\Theta)$;

(ii) any $a \in A_x$ is the limit of a series of points $\{a_n\}$ in the convex hull of $h_x(\Theta)$; since $\tilde{g}$ is linear (hence continuous), $\tilde{g}(a)$ is the limit of the points $\{\tilde{g}(a_n)\}$ in the convex hull of $h_{g(x)}(\Theta)$ and therefore is in $A_{g(x)}$. □

It is easily seen from definition formula (3.5) that $|h_x(\theta)| \leq |h(\theta)|$ holds for all $x \in X$ and all $\theta \in \Theta$; indeed,

$$|\Sigma\alpha(x|\bar{g}_i(\theta))\tilde{g}_i h(\theta)|^2 \leq \Sigma\alpha^2(x|\bar{g}_i(\theta))|\tilde{g}_i h(\theta)|^2$$

$$= |h(\theta)|^2 \Sigma\alpha^2(x|\bar{g}_i(\theta)) \leq |h(\theta)|^2$$

so that h_x leads to a contraction of the space $h(\Theta)$.

Finally, two rather general results are presented on projections. Recall that in a metric space R, the <u>(perpendicular) projection</u> of $x \in R$ on a closed convex subset $S \subset R$ is the (unique) $x_0 \in S$ having the smallest distance to x.

<u>Lemma 3.7</u>. Let $S \subset \mathbb{R}_m$ be closed and convex; let x_0 denote the projection of $x \in \mathbb{R}_m$ on S. Then

$$(3.8) \qquad (x-x_0)^T(x+x_0-2s) > 0, \; x \notin S, \; s \in S$$

<u>Proof</u>. The convexity of S implies

$$|(x-x_0) + (x_0-s)|^2 = |x - s|^2 > |x_0 - s|^2$$

Expanding the left-hand square gives

$$|x - x_0|^2 + 2(x-x_0)^T(x_0-s) > 0$$

implying (3.8). □

<u>Lemma 3.8</u>. Let $S \subset \mathbb{R}_m$ be closed and convex; let x_0 denote the projection of $x \in \mathbb{R}_m$ on S. If $\tilde{g} : \mathbb{R}_m \to \mathbb{R}_m$ is linear and isometric, the projection of $\tilde{g}(x)$ on $\tilde{g}(S)$ is given by $\tilde{g}(x_0)$.

Proof. Denote the projection of $\tilde{g}(x)$ on $\tilde{g}(S)$ by y, so that the distance between $\tilde{g}(x)$ and $\tilde{g}(S)$ equals $|\tilde{g}(x) - y|$; since $\tilde{g}$ is an isometry, this distance is equal to the distance $|x - x_0|$ between x and S. Hence, by the properties of $\tilde{g}$

$$|\tilde{g}(x) - y| = |x - x_0| = |\tilde{g}(x-x_0)| = |\tilde{g}(x) - \tilde{g}(x_0)|$$

and the uniqueness of the projection shows that $y = \tilde{g}(x_0)$. □

Now, all ingredients are collected to cook the main Theorem 3.9.

3.3. Inadmissibility of boundary rules

Theorem 3.9. Consider linvariant quadratic estimation problem (Θ,L,X) and let D_L denote the class of nonrandomized linvariant estimators for this problem. Let $d \in D_L$ and assume that a $\theta \in \Theta$ exists for which $\{x : d(x) \notin A_x\}$ has positive probability, where A_x is defined by (3.6). Then d is strictly dominated by $d_0 \in D_L$, where $d_0(x)$ is defined for all $x \in X$ as the projection of d(x) on A_x.

Proof. The risk function of d equals

$$R(\theta,d) = \int_X f(x|\theta)|h(\theta) - d(x)|^2\mu\{dx\}$$

Let m denote the number of elements of G; then (1.6) implies

$$\begin{aligned} mR(\theta,d) &= \Sigma R(\bar{g}_i(\theta),d) \\ &= \int \Sigma f(x|\bar{g}_i(\theta))|h\bar{g}_i(\theta) - d(x)|^2\mu\{dx\} \\ &= \int \Sigma f(x|\bar{g}_i(\theta))[|\tilde{g}_i h(\theta)|^2 + |d(x)|^2 - 2d^T(x)\tilde{g}_i h(\theta)]\mu\{dx\} \\ &= \int [|h(\theta)|^2 + |d(x)|^2 - 2d^T(x)h_x(\theta)]\Sigma f(x|\bar{g}_i(\theta))\mu\{dx\} \end{aligned}$$

in view of Lemma 3.3 and (3.5). Now, $d_0 g(x)$ is the projection of $dg(x)$ on $A_{g(x)}$, or equivalently, the projection of $\tilde{g}d(x)$ on $\tilde{g}(A_x)$ - by virtue of the definition of an invariant rule and by (3.7). By Lemma 3.8 this projection equals $\tilde{g}d_0(x)$, implying that d_0 is invariant too. Hence, the expression for $mR(\theta,d)$ given above holds for d_0 as well. With the definition

$$V := \{x : d(x) \notin A_x\}$$

d and d_0 coincide on $X - V$ and $m[R(\theta,d) - R(\theta,d_0)]$ may be written as

$$\int_V [d(x) - d_0(x)]^T [d(x) + d_0(x) - 2h_x(\theta)]\Sigma f(x|\bar{g}_i(\theta))\mu\{dx\}$$

Application of Lemma 3.7 (with S replaced by A_x) shows the integrand to be positive for all $x \in V$. Hence $R(\theta,d)$ exceeds $R(\theta,d_0)$ whenever $P_\theta(V)$ is positive. Since projections are measurable functions, as well as the functions h_x, both the set V and the function $d_0 : X \to A$ are measurable, completing the proof. □

The conclusion of the theorem is that any linvariant rule with a value outside A_x can be improved by projecting this value on A_x. So for any observation $x \in X$ the action space A can in fact be reduced to A_x. Of course, this result is only of interest if A_x is a strict subset of A - at least for some $x \in X$.

The role of the invariance assumption can be clarified by concentrating on Bayes rules. According to (2.6), for quadratic loss the Bayes rule equals the posterior expectation of $h(\theta)$. This can be a boundary point of the convex action space only if the posterior distribution of θ is degenerated, which in turn implies that the prior distribution must be generated too (and that the observation must be in agreement with that prior). The invariance, however, excludes priors that are degenerated at a boundary point. Note that for absolute loss on the other hand, a Bayes rule is a median of the posterior distribution of $h(\theta)$, which may be a boundary point of A for a nondegenerated prior. This more or less heuristic argument shows that Theorem 3.9 will not hold for an absolute loss function; compare Example 3.24 (ii).

Theorem 3.9 can be elucidated further with the aid of a new general notion: an inadmissible (random) set. A random set in $\mathbb{R}_m$ is defined as a function V_X of a random variable X, which assigns a set

$V_x \subset \mathbb{R}_m$ to any realisation x of X. Note that any (deterministic) set V can be viewed as a random set V_X with X degenerated.

Definition 3.10. Consider the general estimation problem of Definition 1.4; let D_0 be a class of estimators for this problem. Random set V_X in $\mathbb{R}_m$ is called inadmissible under D_0, if any rule $d \in D_0$ with the property

$$P_\theta\{d(X) \in V_X\} > 0$$

for some $\theta \in \Theta$, is strictly dominated by some $d' \in D_0$. □

The definition implies that any rule in D_0 taking values in V_X (with positive probability for some $\theta \in \Theta$), is inadmissible; hence the name. As a first example, the set $\mathbb{R}_m - C\,\{h(\Theta)\}$ may serve: one of the arguments for the choice of the action space in Definition 1.4 was the inadmissibility under D of this set.

Theorem 3.9 now can be reformulated as follows.

Corollary 3.11. Consider a linvariant quadratic estimation problem (Θ,L,X) and let D_L denote the class of nonrandomized linvariant estimators for this problem. Then random set V_X defined by $V_x := A - A_x$, where A_x is defined by (3.6), is inadmissible under D_L. □

3.4. Applications

In this section three examples are presented, regarding nontruncated estimation problems. The first two examples refer to exponential families of probability distributions: normal and beta distributions respectively. Although the results are not very surprising, these examples are treated in some detail, since their truncated counterparts will be considered in the sequel.

Example 3.12. Consider estimation problem (Θ,L,X) where X has normal distribution $N(\theta,1)$, $\Theta = \mathbb{R}$ and $L(\theta,a) = (\theta-a)^2$. This problem is invariant under group $G = \{e,g\}$ where $g(x) = -x$ for all $x \in \mathcal{X}$. It is easily checked that $\bar{G} = \{e,\bar{g}\}$ and $\tilde{G} = \{e,\tilde{g}\}$ with $\bar{g}(\theta) = -\theta$ and $\tilde{g}(a) = -a$. The problem satisfies the conditions of Theorem 3.9; I-invariant rules are characterized by the property $d(-x) = -d(x)$. The functions h_x in (3.5) become

$$h_x(\theta) = \theta\,\frac{f(x|\theta)-f(x|-\theta)}{f(x|\theta)+f(x|-\theta)} = \theta\,\frac{\exp(\theta x)-\exp(-\theta x)}{\exp(\theta x)+\exp(-\theta x)} = \theta\tanh(\theta x)$$

Figure 3.13 shows the typical behavior of h_x for a positive value of observation x; for negative x, the graph is reflected with respect to the horizontal axis.

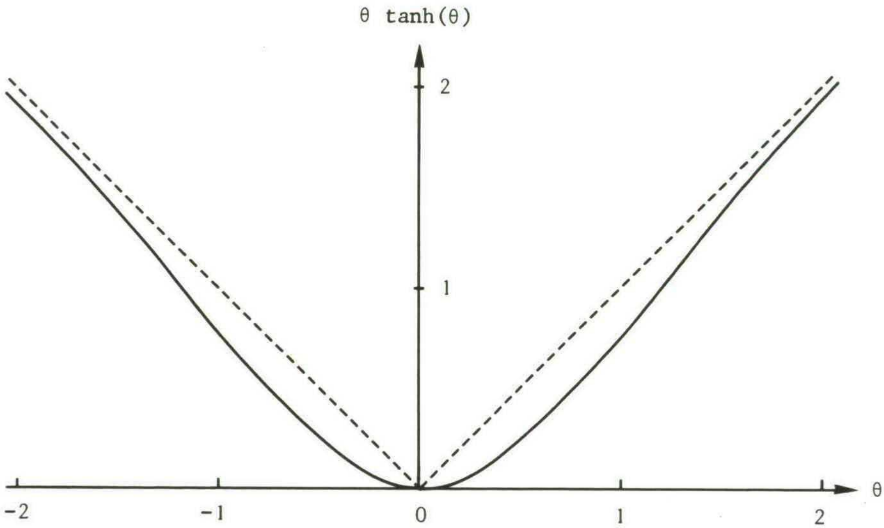


Figure 3.13. The function h_1 in Example 3.12

The space A_x from (3.6) equals $h_x(\mathbb{R})$, hence

$$A_x = \begin{cases} \mathbb{R}_0^+ & \text{for } x > 0 \\ 0 & \text{for } x = 0 \\ \mathbb{R}_0^- & \text{for } x < 0 \end{cases}$$

where $\mathbb{R}_0^+ := [0,\infty)$ and $\mathbb{R}_0^- := (-\infty,0]$. So, for any admissible estimator the signs of observation and estimate have to correspond, which is of course perfectly plausible.

Using Definition 3.10, this result may be formulated equivalently by stating that the set V_X defined by

$$V_x = \begin{cases} \mathbb{R}^+ & \text{for } x < 0 \\ \mathbb{R}^- & \text{for } x > 0 \end{cases}$$

is inadmissible under class D_L; $\mathbb{R}^+$ denotes $(0,\infty)$ and $\mathbb{R}^- := (-\infty,0)$.□

Example 3.14. Consider estimation problem (Θ,L,X), where X has the (two-parametric) beta distribution $Be(\alpha,\beta)$ and $L(\theta,a) = |\theta - a|^2$ with $\theta = (\alpha,\beta)^T \in \Theta = \mathbb{R}^+ \times \mathbb{R}^+$. This problem is invariant under group $G = \{e,g\}$ with g defined by $g(x) = 1-x$; then $\bar{g}(\theta) = (\beta,\alpha)^T$ and

$$\tilde{g} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

So, again a linvariant quadratic estimation problem has been obtained. Now

$$f(x|\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \; 0 < x < 1$$

implies

$$h_x(\theta) = \begin{pmatrix} [1 + y^{\beta-\alpha}]^{-1} & [1 + y^{\alpha-\beta}]^{-1} \\ [1 + y^{\alpha-\beta}]^{-1} & [1 + y^{\beta-\alpha}]^{-1} \end{pmatrix} \theta, \; y := x/(1-x)$$

and A_x is the convex hull of

$$(3.9) \qquad h_x(\Theta) = \left\{\left(\frac{\alpha y^{\alpha-\beta}+\beta}{1+y^{\alpha-\beta}}, \frac{\alpha+\beta y^{\alpha-\beta}}{1+y^{\alpha-\beta}}\right) : (\alpha,\beta) \in \Theta\right\}$$

Figure 3.15 is typical for the way Θ is contracted by the function h_x for $x > 0.5$. The curve in this picture is the image of both axes. The unshaded area is $h_x(\Theta)$; A_x is obtained by taking its closed convex hull and equals the cone between the horizontal axis and line $\alpha = \beta$. In other words, the set V_X defined by

$$V_x = \begin{cases} \{(\alpha,\beta) : 0 < \beta < \alpha\} & \text{for } x < 0.5 \\ \{(\alpha,\beta) : 0 < \alpha = \beta\} & \text{for } x = 0.5 \\ \{(\alpha,\beta) : 0 < \alpha < \beta\} & \text{for } x > 0.5 \end{cases}$$

is inadmissible under the class of linvariant rules. Again, this result is not surprising at all. □

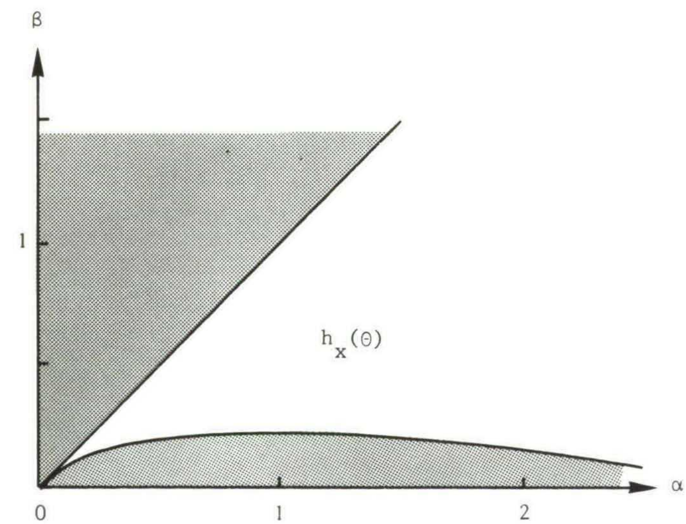


Figure 3.15. Parameter space in Example 3.14, contracted by $h_{0.8}$

(Note that in both examples the parameter space is halved in effect, fully according to intuition.) The third and last example of this section concerns a family of distributions which is not exponential. Here, the parameter space is a finite interval; application of Theorem 3.9 now leads to more interesting findings.

Example 3.16. Consider estimation problem (Θ,L,X), where X has the density

$$f(x|\theta) = \begin{cases} 1 - \theta(\tfrac{1}{2}-x) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Take $L(\theta,a) = (\theta-a)^2$ and $\Theta = [-2,2]$ - the largest interval for which $f(x|\theta)$ is uniformly nonnegative. This problem is linvariant under $G = \{e,g\}$ with $g(x) = 1-x$, $\bar{g}(\theta) = -\theta$ and $\tilde{g}(a) = -a$; hence

$$h_x(\theta) = f(x|\theta)\theta + f(x|-\theta)(-\theta) = (x-\tfrac{1}{2})\theta^2$$

It follows that A_x is the closed interval in $\mathbb{R}$ with endpoints $4x-2$ and 0 for all $x \in X$. Figure 3.17 shows this reduced action space. Note that except for $x = 0$ or $x = 1$, A_x is (much) smaller than half the parameter space and that the action space $A = [-2,2]$ is shortened on both sides. This can be expressed alternatively by calculating the probability that the inadmissible set contains θ. One obtains for $\theta > 0$:

$$P_\theta\{\theta \in V_X\} = P_\theta\{\theta > 4X - 2\} = P_\theta\{X < (\theta+2)/4\}$$

$$= F((\theta+2)/4|\theta) = (\theta^3+4\theta+16)/32$$

where $F(x|\theta) = (1-\theta/2)x + \theta x^2/2$ denotes the distribution function of X. The same result holds for $\theta < 0$, while $P_0\{0 \in V_X\} = 0$. So the conclusion can be drawn that $\theta(\neq 0)$ is in the inadmissible set with a probability of at least $\frac{1}{2}$. The invariance principle therefore leads to a large reduction of the space of possible point estimates. Finally, the MLE is easily seen to equal $2\ \mathrm{sgn}(X-\tfrac{1}{2})$, hence is in the inadmissible set with probability 1 (for $\theta \neq 0$). □

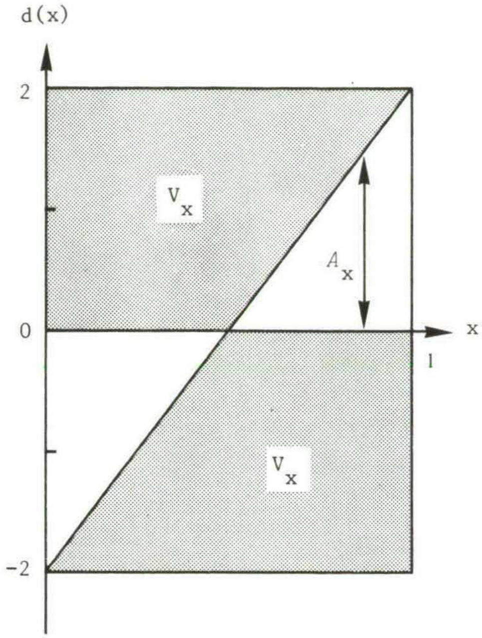


Figure 3.17. Inadmissible set and reduced action space A_x for Example 3.16

3.5. Applications to truncated problems

In this section applications of Theorem 3.9 are presented in the case of truncated estimation problems. As a first example, the very simple problem of Example 2.12 is reconsidered and an inadmissible set derived. Secondly, the truncated analogon of Example 3.14 is treated. The most interesting application shows Example 3.22 where simple inequality constraint regression is considered. Further examples can be found in Chapters 4 and 5. Theorem 4.6 treats the general case of a truncated one-parametric exponential family and is a truncated generalization of Example 3.12. Finally, Section 5.3 presents detailed applications of Theorem 3.9 to several randomized response models.

Example 3.18. Consider again the estimation problem of Example 2.12: $X \in B(1,\theta)$, $L(\theta,a) = (\theta-a)^2$ and $\Theta = (1-P,P)$ with $0.5 < P \leqslant 1$. This problem was seen to be invariant under $G = \{e,g\}$ with $g(x) = 1-x$, while $\bar{g}(\theta) = 1-\theta$ and $\tilde{g}(a) = 1-a$. Theorem 3.9 may be applied in this situation, giving

$$h_0(\theta) = 2\theta(1-\theta)$$

and $A_0 = [2P(1-P),\ 0.5]$. A similar result holds for $x = 1$ and using $\phi := (2P-1)^2$ leads to

$$A_x = \begin{cases} [(1-\phi)/2,\ 0.5] & \text{for } x = 0 \\ [0.5,\ (1+\phi)/2] & \text{for } x = 1 \end{cases}$$

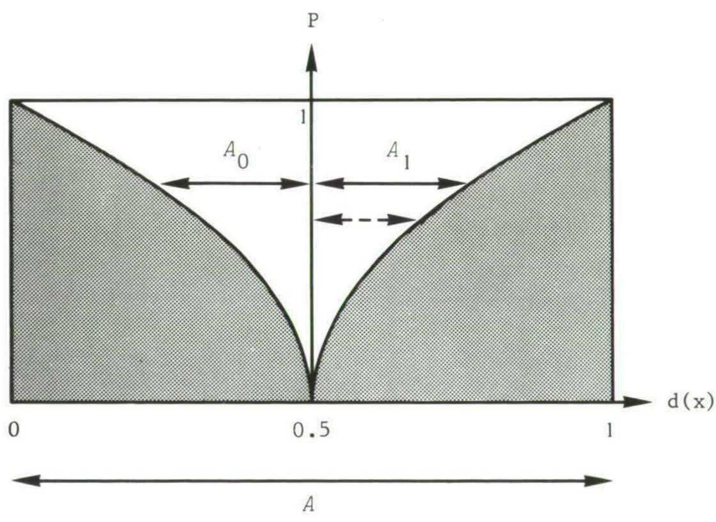


Figure 3.19. Restricted action space A_x in Example 3.18

Figure 3.19 shows this effective action space for a given P. Note that the action space A has been reduced fairly sizable. A practical illustration can be given by borrowing the interpretation of this problem, offered in Example 2.4. The probability of throwing heads with a given coin is known with certainty to be between 0.2 and 0.8. The statistician is asked to estimate this probability after a single toss with the coin

- which produced heads. It follows that any sensible estimate has to be between 0.5 and 0.68 (compare the dotted line segment in Figure 3.19).

Since the estimation problem considered here is very simple, the results could have been obtained more directly. To see this, the risk function of a linvariant (nonrandomized) rule d_p has to be considered, which is given by (2.11). From the partial derivative

$$\frac{\partial R(\theta, d_p)}{\partial p} = 4\theta(1-\theta) - 2(1-p)$$

it follows that for fixed θ the risk is increasing for

$$p > 1 - 2\theta(1-\theta)$$

But then the risk is increasing for all $\theta \in \Theta$ for

$$p > \max_{\theta \in \Theta} \{1 - 2\theta(1-\theta)\} = (1+\phi)/2$$

So, the estimate for θ should never exceed $(1+\phi)/2$, which is in full agreement with the earlier results. Note that for $\phi \leqslant 0.5$ the minimax rule d_m, given by (2.13) exactly coincides with the upper endpoint of A_1, and the lower endpoint of A_0. □

Example 3.20. Consider again the decision problem of Example 3.14. Assume that available prior knowledge states that X is more or less uniformly distributed, formalized in the assumption that the parameters of the distribution $Be(\alpha,\beta)$ are both between 0.5 and 1.5; hence $\Theta = (0.5, 1.5) \times (0.5, 1.5)$. A_x can be found from (3.9); note that h_x maps boundaries of Θ onto the boundaries of $h_x(\Theta)$. In Figure 3.21 $h_x(\Theta)$ is drawn as well as its convex closure A_x for a given value of x. In general, A_x is the triangle with corner points (0.5, 0.5), (1.5, 1.5) and ((3p+1)/(2p+2), (p+3)/(2p+2)). □

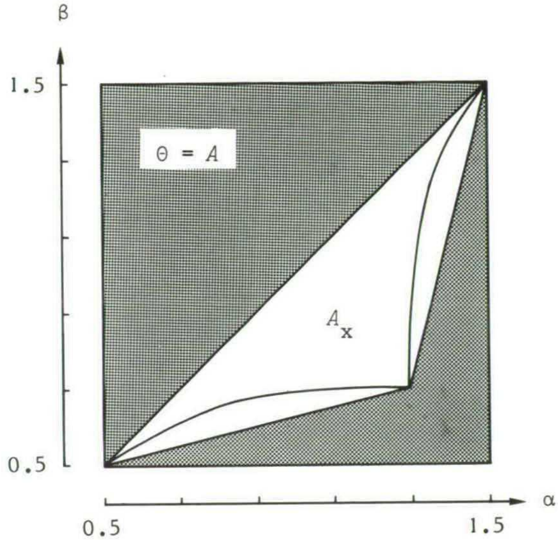


Figure 3.21. Inadmissible set (shaded area) for Example 3.20 with x = 0.8

Example 3.22. Consider the simple regression model

$$Y_i = \theta_1 + \beta z_i + \varepsilon_i \ , \ i = 1,2,\ldots,n$$

with the usual assumptions: z_i nonstochastic, $\varepsilon_i \sim N(0,\sigma^2)$ and $E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$. Assume for simplicity $\sigma^2 = 1$ and let β be allowed to be interpreted as a fraction. E.g., Y_i and z_i denote consumption and income respectively of individuals or households; then β is the (marginal) consumption quote and should satisfy $0 \leqslant \beta \leqslant 1$. The model can be rewritten for simplicity as

$$X_i = \theta_1 + \theta_2 z_i + \varepsilon_i$$

where $X_i := Y_i - z_i/2$ and $\theta_2 := \beta - \frac{1}{2}$. The problem of how to estimate $\theta := (\theta_1, \theta_2)^T$ with $\Theta = \mathbb{R} \times [-\frac{1}{2}, \frac{1}{2}]$ is linvariant with $g(x) = -x$, $\bar{g}(\theta) = -\theta$ and $\tilde{g}(a) = -a$. Application of Theorem 3.9 gives:

$$f(x|\theta)/f(x|\bar{g}(\theta)) = \exp[2 \sum_i x_i(\theta_1+\theta_2 z_i)]$$

$$h_x(\theta) = \theta \tanh[\sum_i x_i(\theta_1+\theta_2 z_i)].$$

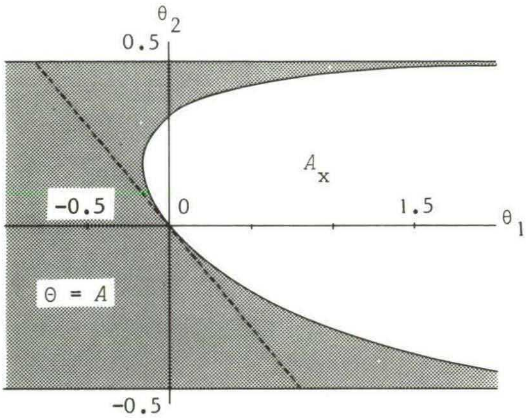


Figure 3.23. Inadmissible set V_x (shaded area) for Example 3.22 with $\Sigma x_i = 1$ and $\Sigma x_i z_i = 1.6$

Figure 3.23 presents the set A_x in case $\Sigma_i x_i z_i$ and $\Sigma_i x_i$ are both positive. It can be obtained easily by observing that h_x maps any line through the origin onto itself. Again A_x turns out to be a strict subset of Θ; the usual estimators are not admissible. It is easy to check that for $\theta_1 = 0$ the inadmissible set for regression without constant term is obtained. □

It will be very interesting to extend Example 3.22 to more complex cases, with more independent variables, more equations and unknown variances. For example, the general linear model

$$Y = Z\beta + \varepsilon$$

where the observations consist of the deterministic n×k-matrix Z and the stochastic n-vector Y, can be analyzed in the same way. Under the given linear constraints

$$a \leqslant A\beta \leqslant b$$

the action (parameter) space $\{\beta : a \leqslant A\beta \leqslant b\}$ can be restricted to a smaller subset, which depends upon the observations. Shifting the origin towards a point of symmetry of the parameter space again simplifies matters considerably. Since inequality constraint regression is a popular topic (see TOUTENBURG 1982 for a general treatment, or more specificly LIEW 1976, DAVIS 1978 or CHANG 1981), further research in this direction will be of importance.

3.6. Extensions

The general scope of Theorem 3.9 is that dominating estimators are found by a contraction method. This is a feature common to many other estimators in similar situations. Compare the shrunken estimators and ridge estimators in regression, and the general Hunt-Stein type of estimators.

The findings in this chapter may be generalized and extended in several directions. Perhaps the most obvious generalization is to use instead of (1.3) the weighted quadratic loss function

$$(3.10) \qquad L(\theta,a) = w(\theta)\,|h(\theta) - a|^2$$

where w is continuous and positive on Θ. Under the additional invariance assumption

$$(3.11) \qquad w(\theta) = w\bar{g}(\theta) \ , \ \theta \in \Theta$$

for any $\bar{g} \in \bar{G}$, it is easy to check that Lemma 3.3 remains valid, as well as Theorem 3.9.

It is conjectured, however, that similar results can be obtained for any regular estimation problem with strictly convex loss function. This will be illustrated by reconsidering Example 3.18, now using different loss functions.

Example 3.24. Consider estimation problem (Θ, L, X), where $X \in B(1,\theta)$ and $\Theta = (1-P,P)$ with $\frac{1}{2} < P < 1$. Three loss functions will be considered, all of them keeping the problem linvariant with $\tilde{g}(a) = 1-a$. Note that A is bounded, rendering condition (ii) of Definition 1.5 irrelevant.

$$\text{(i)}\ L(\theta,a) = \frac{(\theta-a)^2}{\theta(1-\theta)}$$

As was noted in Example 1.14, this loss function is of interest, since in the case $P = 1$, X itself is a minimax estimator. The loss is of type (3.10) where the weight function $w(\theta) = [\theta(1-\theta)]^{-1}$ has the invariance property (3.11); it follows that Theorem 3.9 holds. Hence, for this strictly convex loss function boundary rules will be inadmissible and the action space can be reduced, just as in Example 3.18. Note that the alternative derivation at the end of Example 3.18 remains valid as well.

$$\text{(ii)}\ L(\theta,a) = |\theta - a|$$

Now the risk function satisfies for d_p defined by $d_p(1) = p$

$$R(\theta,d_p) = \theta|\theta - p| + (1-\theta)|\theta - (1-p)|$$

$$R(P,d_p) = P(P-p) + (1-P)(P-(1-p)) = (2P-1)(1-p)$$

Hence, for $\theta = P$, the risk is minimized by giving p its maximum value P, so that estimates on the boundary of A not necessarily correspond to an inadmissible estimator in the case of this convex (but not strictly) loss function.

$$\text{(iii)}\ L(\theta,a) = (1-\theta)a + \theta(1-a)$$

Here, the risk function

$$R(\theta,d_p) = \theta[(1-\theta)p + \theta(1-p)] + (1-\theta)[(1-\theta)(1-p) + \theta p]$$

$$= -p(2\theta-1)^2 + \theta^2 + (1-\theta)^2$$

is minimized uniformly in θ by giving p its maximum value P. Again, admissible estimators may take values on the boundary of A for this (not strictly) convex loss function. □

Another direction in which this chapter could be generalized can be found by considering estimation problems that can be truncated to invariant problems, although they are not invariant themselves.

Definition 3.25. An estimation problem (Θ,L,X) is called pseudo-invariant, if an open set $\Theta_0 \subset \Theta$ exists such that problem (Θ_0,L,X) is invariant. □

As an application, note that Example 2.6 resulted in inadmissible boundary rules if and only if $Q < \frac{1}{2} < P$, that is, if and only if the estimation problem considered is pseudo-invariant.

Finally, it is conjectured that the approach of this chapter can be extended to include infinite, but compact groups G. (An example of such a group is the group consisting of all rotations of the plane.)

4. EXPONENTIAL FAMILIES

4.1. Introduction and summary

In this chapter regular exponential families of probability distributions will be discussed, which are given in the canonical form. Hence, for a one-parametric exponential family the densities are found by specializing (1.13) to

$$(4.1) \qquad f(x|\theta) = \exp[\theta t(x) + a(x) - c(\theta)] \ , \ x \in X$$

See Section 1.5 for a summary of the most important properties of exponential families. The investigations in this chapter will extend into two directions.

In the next section exponential families will be considered that are invariant under a group G of measure preserving functions of X onto itself; the decision problem concerns the estimation of parameter θ itself. After some preliminary lemmas it is shown that in this setting all invariant estimation problems involving one-parametric exponential families are automatically linvariant. It follows that Theorem 3.9 is applicable; in the special case that G consists of two elements, a general expression is derived for the reduced action space A_x. This result generalizes some of the examples of the previous chapter. To conclude Section 4.2, the case that the estimand θ is a two-dimensional vector is discussed briefly. A survey is given of all possible ways in which such estimation problems may be linvariant under a group of two elements.

Section 4.3 starts the discussion of Bayes estimators in the case of a one-parametric exponential family. Conjugated families of prior distributions are considered and the emphasis lies on truncated regular problems. (Generalized) Bayes estimators with finite Bayes risk are shown to be admissible. The estimand not necessarily is θ itself; an important application concerns the situation that $h(\theta)$ is the expectation of the sufficient statistic $t(X)$. Although for infinite Θ the generalized Bayes rule with respect to the uniform measure may have infinite Bayes risk, this rule is admissible in this case (KATZ 1961).

Section 4.4 considers the special case of an (exponential) family of exponential distributions with truncated parameter space. Both for the expectation and the variance a class of admissible estimators is derived. These estimators appear to be closely related to Mills' ratio $\Phi(-x)/\phi(x)$, where Φ and ϕ denote the distribution function and the density respectively of the standard normal distribution. Quite a number of inequalities were derived for this ratio, e.g. FELLER 1950, p. 193. As a side-product of the analysis in this section, the Feller bounds can be proved very simply and sharper bounds are obtained.

4.2. Invariant exponential families

To simplify the proof of the main theorem of this section three simple lemmas are given first. They concern arbitrary functions $u_i : S \to \mathbb{R}$ and $v_i : T \to \mathbb{R}$ $(i = 1,2,3)$. By way of example Lemma 4.3 is proved; the proofs of the other two lemmas are omitted.

Lemma 4.1. Assume that

$$u_1(x)v_1(y) = u_2(x), \quad x \in S, \ y \in T$$

Then either u_1 is identically zero on S or v_1 is constant on T. $\square$

Lemma 4.2. Assume that

$$(4.2) \qquad u_1(x)v_1(y) = u_2(x) + v_2(y), \quad x \in S, \ y \in T$$

Then either u_1 is constant on S or v_1 is constant on T. $\square$

Lemma 4.3. Assume that

$$(4.3) \qquad yu_1(x) + u_2(x)v_1(y) = u_3(x), \quad x \in S, \ y \in T$$

where T contains at least two points. Then a $p \in \mathbb{R}$ exists such that

$$u_1(x) = pu_2(x), \quad x \in S$$

Proof. Let $a \in T$, $b \in T$, $a \neq b$; substituting these in (4.3) and subtracting the resulting equations gives

$$(a-b)u_1(x) + [v_1(a) - v_1(b)]u_2(x) = 0, \quad x \in S$$

Take $p = [v_1(a) - v_1(b)]/(b-a)$ to complete the proof. □

Now consider the exponential family with densities (4.1), which is invariant under a group G. Again, $\bar{G}$ is the group of corresponding functions: $\bar{g} : \Theta \to \Theta$. It can be expected that all $\bar{g}$ are linear, since θ occurs linearly in the cross product of (4.1). The next theorem shows that this conjecture is basically correct.

Theorem 4.4. Assume that the regular one-parametric exponential family with density (4.1) is invariant under a group G of measure preserving functions g. Then all functions $\bar{g}$ from the corresponding group $\bar{G}$ are linear.

Proof. Under the assumptions stated, Lemma 3.4 is applicable, hence (3.4) holds almost everywhere for fixed θ and $\bar{g}$. In combination with (4.1) this gives

$$\bar{g}(\theta)t(x) + a(x) - c\bar{g}(\theta) = \theta t g^{-1}(x) + a g^{-1}(x) - c(\theta) \tag{4.4}$$

for all $g \in G$, all $\theta \in \Theta$ and almost all $x \in \mathcal{X}$. This remains true when x is replaced by g(x):

$$\bar{g}(\theta)tg(x) + ag(x) - c\bar{g}(\theta) = \theta t(x) + a(x) - c(\theta)$$

Subtracting this relation from (4.4) leads to

$$\theta[tg^{-1}(x) - t(x)] + \bar{g}(\theta)[tg(x) - t(x)] = 2a(x) - ag(x) - ag^{-1}(x) \tag{4.5}$$

again holding for all $g \in G$, all $\theta \in \Theta$ and almost all $x \in \mathcal{X}$.

Application of Lemma 4.3 to (4.5) shows that for all $g \in G$ a $p_g \in \mathbb{R}$ exists with

$$(4.6) \qquad tg^{-1}(x) - t(x) = p_g[tg(x) - t(x)]$$

a.e. Substitution of (4.6) into (4.5) gives

$$(4.7) \qquad [\bar{g}(\theta) + p_g\theta][tg(x) - t(x)] = 2a(x) - ag(x) - ag^{-1}(x)$$

Application of Lemma 4.1 now leads for any $g \in G$ to two possibilities:

(i) The first factor on the left-hand side of (4.7) is constant, implying that for some $q_g \in \mathbb{R}$

$$(4.8) \qquad \bar{g}(\theta) = q_g - p_g\theta, \ \theta \in \Theta$$

(ii) The second factor on the left-hand side of (4.7) is zero, hence $tg = t$ a.e. and $tg^{-1} = t$ by (4.6). Substituting this in (4.4) gives

$$(4.9) \qquad [\theta - \bar{g}(\theta)]t(x) = a(x) - ag^{-1}(x) + c(\theta) - c\bar{g}(\theta)$$

Finally, apply Lemma 4.2 to (4.9) to see that one of the left-hand side factors is constant. Since t is not constant on X by condition (ii) of Definition 1.16, a $q_g \in \mathbb{R}$ exists such that $\bar{g}(\theta) = q_g + \theta$, $\theta \in \Theta$. However, this relation is a special case of (4.8), so that the general conclusion can be drawn that (4.8) holds for all $g \in G$. □

This result can be applied to invariant estimation problems (Θ,L,X) with $L(\theta,a) = (\theta-a)^2$, where X has a density (4.1). If G consists of measure preserving functions, it follows that the problem automatically is linvariant; next, Theorem 3.9 will be used to obtain a general expression for the reduced action space A_x or equivalently, for the functions h_x, in the case $G = \{e,g\}$.

Theorem 4.5. Consider estimation problem (Θ,L,X) with $L(\theta,a) = (\theta-a)^2$ and regular exponential family $\{P_\theta : \theta \in \Theta\}$ with density (4.1). Assume that the problem is invariant under a group G of measure preserving functions with corresponding groups $\bar{G}$ and $\tilde{G}$. Then $\bar{G}$ (as well as $\tilde{G}$) consists of translations and reflexions only.

Proof. Since h is the identity, Lemma 3.3 (i) implies that groups $\bar{G}$ and $\tilde{G}$ are isomorphic, hence all $\bar{g} \in \bar{G}$ are isometric as well as linear (by virtue of Theorem 4.4). Hence for all $g \in G$, p_g in (4.8) must have absolute value 1, so that either

$$\bar{g}(\theta) = \theta + q_g, \ \theta \in \Theta$$

must hold (representing a translation), or

$$\bar{g}(\theta) = -\theta + q_g, \ \theta \in \Theta$$

(a reflexion with respect to $q_g/2$). □

Theorem 4.6. Consider estimation problem (Θ,L,X) with $L(\theta,a) = (\theta-a)^2$ and regular exponential family $\{P_\theta : \theta \in \Theta\}$ with density (4.1). Assume that the problem is invariant under a group of exactly two measure preserving functions. Then the reduced action space A_x can be derived from the function

$$h_x(\theta) = \theta \tanh[\theta t(x)] \tag{4.10}$$

Proof. Following Theorem 4.5, group G must contain identity e and a reflexion. The latter may be taken to be the reflexion with respect to the origin; because of the indefiniteness in (4.1) - compare the remark preceding Example 1.19 - this implies no loss of generality. Substitution of $\bar{g}(\theta) = -\theta$ into (4.5) implies ag = a on X, while (4.4) can be rewritten as

$$\theta[tg(x) + t(x)] = c(\theta) - c\bar{g}(\theta)$$

According to Lemma 4.1, $tg(x) + t(x)$ then is a constant for all $x \in X$, for which the value zero can be chosen - again without loss of generality. Hence $tg = -t$ on X and $c\bar{g} = c$ on Θ. Substitution into (3.5) gives the desired result. □

Example 4.7. Let the assumptions of Theorem 4.6 hold and suppose that the observations lead to the value $t(x) = 1$. The behavior of h_x is shown by Figure 3.13; indeed, Example 3.12 is a special case of the above theorem. Assume further that $\Theta = (-m,m)$. Then the reduced action space A_x is the interval $[0, m \tanh m]$.

More explicitly, let the exponential family consist of the normal distributions $N(\theta,\sigma^2)$ with known σ^2; so, one may take $t(x) = x/\sigma^2$ in (4.1) - compare (1.17). Assume that $\Theta = (-1,1)$. Then the interval $[0, 0.762]$ is the reduced action space corresponding with the 'observation' $t(x) = 1$. Hence, even if x happens to equal σ^2, estimates exceeding 0.762 correspond with inadmissible estimators - whatever the value of σ^2 may be. □

To conclude this section some attention is paid to linvariant decision problems, where the estimand θ is a vector. Assume that the problem is linvariant under a group of exactly two measure preserving functions, say e and g. Then the corresponding linear function $\tilde{g} : A \to A$ must be isometric and tripotent. (The latter property means that $\tilde{g}\tilde{g}\tilde{g} = \tilde{g}$ and holds, since $\tilde{G}$ is a group.) Because of the linearity, $\tilde{g}$ can be denoted by a square matrix M; the isometry implies

$$(a,a) = (Ma,Ma) = (M^T Ma,a)$$

for all $a \in A$. Since A contains an open set in $\mathbb{R}_k$, $M^T M = I$. Combining this with the group property $M^2 = I$ shows that M is symmetric.

In the two-parametric case all possible isometries of the plane are rotations and reflexions, which can be represented by the matrices

$$\begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}, \begin{pmatrix} \cos\phi & \sin\phi \\ \sin\phi & -\cos\phi \end{pmatrix}$$

respectively, with $0 \leq \phi < 2\pi$. Reflexions are always symmetric, rotations only for $\sin\phi = 0$, leading to the matrices I and -I. It follows that for these linvariant two-dimensional estimation problems $\tilde{g}$ is one of the matrices

$$(4.11) \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}; \begin{pmatrix} \cos\phi & \sin\phi \\ \sin\phi & -\cos\phi \end{pmatrix}, \ 0 \leq \phi < 2\pi$$

Note that Example 3.14 (as well as 3.20) concerned a reflexion with $\phi = \pi/2$.

4.3. Bayes estimators

In this section the regular estimation problem (Θ,L,X) will be discussed, where the probability distribution of X belongs to a regular one-parametric exponential family and $L(\theta,a) = [h(\theta) - a]^2$. The parameter space $\Theta \subset \mathbb{R}$ will be a finite or infinite open interval (θ_0,θ_1). Contrary to the preceding section, no invariance will be assumed. For this problem Bayes rules will be derived with special attention for admissibility.

For the exponential family determined by (4.1), a conjugated family has densities

$$(4.12) \quad \tau_\alpha(\theta) = \exp[\theta\alpha_1 + v(\alpha) - \alpha_2 c(\theta)], \ \theta \in \Theta$$

according to (2.8); it will be assumed now that these are densities with respect to the Lebesgue measure. As announced in Section 2.4 the space A of parameter values $\alpha := (\alpha_1,\alpha_2)$ will be assumed to contain the vector 0, corresponding to the uniform prior; hence for infinite Θ, τ_0 is only a measure.

The posterior distribution $\tau_\alpha(\theta|x)$ belongs to the same family; it follows from the joint density

$$f_\alpha(x,\theta) = \exp[\{t(x) + \alpha_1\}\theta + a(x) + v(\alpha) - (1+\alpha_2)c(\theta)]$$

As follows from (2.7),

$$(4.13) \quad d_\alpha(x) = \int_{\theta_0}^{\theta_1} h(\theta) f_\alpha(x,\theta)d\theta \bigg/ \int_{\theta_0}^{\theta_1} f_\alpha(x,\theta)d\theta, \ x \in \mathcal{X}$$

is (generalized) Bayes with respect to τ_α. The admissibility of the rules (4.13) is the subject of the next theorem.

<u>Theorem 4.8</u>. Consider problem (Θ,L,X), where Θ is the (possibly infinite) interval (θ_0,θ_1), X has density (4.1) and $L(\theta,a) = [h(\theta) - a]^2$; assume that h is continuous. If $r(\tau_\alpha,d_\alpha)$ is finite, the estimator d_α in (4.13) is (generalized) Bayes with respect to τ_α, $\alpha \in A$ and admissible.
<u>Proof</u>. That d_α is (generalized) Bayes follows by definition. The risk function is continuous, since the conditions of Theorem 1.20 are satisfied: the continuity of h implies condition (ii) and condition (i) is satisfied with $B_1 = 1$ and $B_2 = [h(\theta_1) - h(\theta_2)]^2$; note that B_2 is bounded on compact sets of $\Theta \times \Theta$, since h is continuous. Now Theorem 1.9 proves the admissibility of d_α. □

The case that the decision problem concerns estimation of the derivative $c'(\theta)$ of $c(\theta)$ was previously considered by KARLIN 1958 and KATZ 1961. This choice has some mathematical conveniences; besides, the function $c'(\theta)$ is of great interest, since it equals the expectation of the complete sufficient statistic $t(X)$; see (1.14). So, attention will now be concentrated on the special estimand

$$(4.14) \qquad h(\theta) = c'(\theta) = E_\theta t(X), \quad \theta \in \Theta$$

Since $c(\theta)$ is (infinitely often) differentiable, this function is continuous. If Θ is finite or if τ_α is a proper probability distribution, it is not difficult to show that the (minimum) Bayes risk is finite, so that Theorem 4.8 is applicable. The case $\alpha = 0$ presents more difficulties and will be considered now in detail. It is of special interest, because the (generalized) Bayes rule with respect to the uniform prior τ_0 has a simple form and is often minimax. To stress the differences the case of finite Θ is considered separately.

<u>Lemma 4.9</u>. Consider problem (Θ,L,X), where Θ is the finite interval (θ_0,θ_1), X has density (4.1) and $L(\theta,a) = [c'(\theta) - a]^2$. Then the Bayes estimator d_0 with respect to the uniform prior τ_0 is given by

$$d_0(x) = t(x) + G_0^{-1}(x) - G_1^{-1}(x)$$

$$G_i(x) := \int_{\theta_0}^{\theta_1} \exp[(\theta-\theta_i)t(x) - c(\theta) + c(\theta_i)]d\theta, \; x \in X, \; i = 0,1$$

and d_0 is admissible for $c'(\theta)$, $\theta \in \Theta$.

Proof. By (4.13) the Bayes rule d_0 with respect to τ_0 is given by

$$d_0(x) = \int_{\theta_0}^{\theta_1} c'(\theta) \exp[\theta t(x) - c(\theta)]d\theta \Bigg/ \int_{\theta_0}^{\theta_1} \exp[\theta t(x) - c(\theta)]d\theta$$

The numerator on the right equals

$$-\int_{\theta_0}^{\theta_1} [\exp\{-c(\theta)\}]' \exp[\theta t(x)]d\theta = t(x) \int_{\theta_0}^{\theta_1} \exp[\theta t(x) - c(\theta)]d\theta$$

$$+ \exp[\theta_0 t(x) - c(\theta_0)] - \exp[\theta_1 t(x) - c(\theta_1)]$$

by partial integration, leading to the desired form of $d_0(x)$. With

$$H(x) := G_0^{-1}(x) - G_1^{-1}(x)$$

it follows

$$R(\theta,d_0) = E_\theta[t(X) + H(X) - c'(\theta)]^2$$

$$= V_\theta t(X) + 2 \operatorname{Cov}_\theta[t(X), H(X)] + E_\theta H^2(X)$$

As all terms are finite, the Bayes risk of d_0 is finite as well; Theorem 4.8 completes the proof. □

If the parameter space equals the infinite interval (θ_0,∞), τ_0 is only a measure; hence, the Bayes risk $r(\tau_0,d_0)$ can be infinite, so that Theorem 4.8 is not applicable. Nevertheless, it can be shown that even in this case the generalized Bayes rule d_0 is admissible. The notation

$$(4.15)\quad G(x) := \int_{\theta_0}^{\infty} \exp[(\theta-\theta_0)t(x) - c(\theta) + c(\theta_0)]d\theta,\ x \in X$$

will be useful.

Theorem 4.10. Consider problem (Θ,L,X), where $\Theta = (\theta_0,\infty)$, X has density (4.1) and $L(\theta,a) = [c'(\theta) - a]^2$. Then the generalized Bayes estimator with respect to the uniform measure τ_0 is given by

$$(4.16)\quad d_0(x) = t(x) + G^{-1}(x),\ x \in X$$

and d_0 is admissible for $c'(\theta)$, $\theta \in \Theta$. □

This result is due to KATZ 1961 (covered also by ZACKS 1971). Expression (4.16) follows by the same method as in the foregoing proof. To prove admissibility, d_0 is considered as limit of a series of Bayes rules with respect to proper probability distributions; a detailed comparison of risk functions is needed.

Note that the estimator defined by (4.16) is biased for $c'(\theta)$, $\theta \in \Theta$, in agreement with Theorem 2.3.

Two examples, taken from KATZ 1961, will serve to illustrate Theorem 4.10.

Example 4.11. The density of the normal distribution $N(\theta,1)$ was given in (1.17); let $\Theta = \mathbb{R}^+$. Now (4.15) becomes

$$G(x) = \int_0^{\infty} \exp[\theta x - \theta^2/2]d\theta,\ x \in \mathbb{R}$$

and (4.16) reduces to

$$d_0(x) = x + \phi(x)/\Phi(x),\ x \in \mathbb{R}$$

Here, ϕ and Φ denote density and distribution function respectively of the standard normal distribution. With the use of Mills' ratio $R(x)$, defined by

(4.17) $\quad R(x) := \Phi(-x)/\phi(x),\ x \in \mathbb{R}$

(MILLS 1926), Theorem 4.10 leads to the conclusion that

(4.18) $\quad d_0(x) = x + R^{-1}(-x),\ x \in \mathbb{R}$

is admissible for θ, $\theta \in \mathbb{R}^+$. KATZ 1961 further shows that this estimator is minimax as well. □

Mills' ratio has found quite a number of rather different applications and has been investigated thoroughly. In particular, there has been considerable interest in the derivation of upper and lower bounds for $R(x)$. See JOHNSON & KOTZ 1970, p. 278 ff. for a general discussion.

Example 4.12. The density of the binomial distribution $B(n,\theta)$ was given by (1.16) with $\pi = \log[\theta/(1-\theta)]$. Comparison with (1.13) shows that $c(\pi) = n \log(1+e^{\pi})$, hence by (4.14)

$$h(\pi) = c'(\pi) = ne^{\pi}/(1+e^{\pi}) = n\theta$$

For $\Pi = (\pi_0,\infty)$, (4.15) now reads

$$G(x) = \int_{\pi_0}^{\infty} \exp[(\pi-\pi_0)x - n\log(1+e^{\pi}) + n\log(1+e^{\pi_0})]\,d\pi$$

$$= e^{-\pi_0 x}(1+e^{\pi_0})^n \int_{\pi_0}^{\infty} e^{\pi x}(1+e^{\pi})^{-n} d\pi$$

Introduction of $Q := e^{\pi_0}/(1+e^{\pi_0})$ and change of variables gives

$$G(x) = Q^{-x}(1-Q)^{x-n} \int_Q^1 \theta^x(1-\theta)^{n-x} \frac{d\theta}{\theta(1-\theta)}\ ,\ x = 0,1,\dots,n$$

Theorem 4.10 then implies that

$$(4.19)\quad d_0(x) = \frac{x}{n} + \frac{Q^x(1-Q)^{n-x}}{n\int_Q^1 \theta^{x-1}(1-\theta)^{n-x-1}d\theta}, \quad x = 0,1,\ldots,n$$

is admissible for θ, $\theta \in (Q,1]$. (Note that $d_0(n)$ is on the boundary of the parameter space; yet, the admissibility is not in contradiction with Example 2.4, since the present problem is truncated one-sidedly.) It is not difficult to show that here d_0 is not minimax in general. □

In Section 4.4 a more elaborate application of the theory discussed in this section will be presented for the case of a truncated exponential distribution. Admissible estimators for expectation and variance will be derived; as a by-product, new bounds for Mills' ratio will be found.

4.4. Truncated exponential distributions and Mills' ratio

In this section the following decision problem (Θ,L,X) will be considered. The random variable X has exponential distribution Ne(θ) with density

$$(4.20)\quad f(x|\theta) = \exp[-\theta x + \log\theta], \; x \in \mathbb{R}^+$$

Parameter space Θ is assumed to have a one-sided truncation: $\theta > \theta_0 > 0$ with θ_0 given. Without loss of generality θ_0 may be assumed to be 1, so that $\Theta = (1,\infty)$. Prior distributions will be considered that belong to a conjugated family; more specifically

$$(4.21)\quad \tau_\alpha(\theta) \propto \theta^\alpha, \; \theta > 1$$

Only for $\alpha < -1$, this represents a proper density; however, $A = (-\infty,0]$ will be chosen, so that again the uniform prior measure is included.

In agreement with (4.14), the estimand $h(\theta) = 1/\theta$ is considered, being the expectation of density (4.20). Introducing

$$(4.22)\quad J_\alpha(x) := \int_1^\infty \theta^\alpha \exp[(1-\theta)x]d\theta, \; x \in \mathbb{R}^+, \; \alpha \in \mathbb{R}$$

(4.13) gives the corresponding Bayes rules for $1/\theta$, $\theta \in \Theta$:

$$(4.23) \qquad d_\alpha(x) = J_\alpha(x)/J_{\alpha+1}(x), \quad x \in \mathbb{R}^+$$

The preceding section immediately shows that d_α is admissible for all $\alpha < -1$. It is intuitively clear that d_α should be increasing in x with range [0,1]. Indeed, this follows formally from the next theorem, that lists some properties of J_α.

Theorem 4.13. The function J_α defined by (4.22) satisfies:

(i) $xJ_{\alpha+1}(x) = (\alpha+1)J_\alpha(x) + 1$

(ii) $J_\alpha(x)$ is strictly decreasing in x and strictly increasing in α

(iii) $J^2_\alpha(x) < J_{\alpha-1}(x)J_{\alpha+1}(x)$

(iv) $J'_\alpha(x) = J_\alpha(x) - J_{\alpha+1}(x)$

(v) $[J_\alpha(x)/J_{\alpha+1}(x)]' > 0$

for all $\alpha \in \mathbb{R}$ and all $x > 0$.

Proof. Property (i) follows by partial integration of

$$(\alpha+1)J_\alpha(x) = \int_1^\infty [\theta^{\alpha+1}]' \exp[(1-\theta)x]d\theta$$

while (ii) is a result of the monotonous character of the integrand. Property (iii) is a direct application of the Cauchy-Schwarz inequality. Differentiation under the integral sign gives (iv), and (v) is a immediate consequence of (iii) and (iv). Note that the derivatives of all orders are uniformly positive, implying a.o. that J_α is strictly convex on $\mathbb{R}^+$ for all $\alpha \in \mathbb{R}$. □

Since, from Theorem 4.13 (i),

$$J_0(x) = 1/x \ , \ J_1(x) = (1+x)/x^2$$

the special case $\alpha = 0$ leads to the estimator

(4.24) $\quad d_0(x) = x/(x+1)$

which is admissible for $1/\theta$, $\theta \in \Theta$ by virtue of Theorem 4.10. Of course, this rule follows from (4.13) equally simply.

By a similar approach admissible Bayes estimators can be found for the variance $1/\theta^2$ of exponential density (4.20). The Bayes estimator d_α^* for $1/\theta^2$, $\theta \in \Theta$ with respect to conjugated prior (4.21) is

(4.25) $\quad d_\alpha^*(x) = J_{\alpha-1}(x)/J_{\alpha+1}(x), \ x \in \mathbb{R}^+$

It is easily checked that d_α^* is admissible for all $\alpha < -1$; it is conjectured that admissibility holds for $\alpha = 0$ as well.

The foregoing admissibility results are summarized below.

Lemma 4.14. Consider truncated decision problem (Θ,L,X) with $X \in Ne(\theta)$, $\Theta = (1,\infty)$ and quadratic loss. Then the estimators d_α and d_α^* given by

(4.23) $\quad d_\alpha(x) = J_\alpha(x)/J_{\alpha+1}(x), \ x \in \mathbb{R}^+$

and

(4.25) $\quad d_\alpha^*(x) = J_{\alpha-1}(x)/J_{\alpha+1}(x), \ x \in \mathbb{R}^+$

where $J_\alpha(x)$ is defined by (4.22), are admissible for the expectation $1/\theta$ and the variance $1/\theta^2$ respectively for $\alpha < -1$. Besides, d_α is admissible for $\alpha = 0$. $\square$

The functions J_α are related to Mills' ratio, as the next derivation illustrates.

$$J_{-\frac{1}{2}}(x) = e^x \int_1^\infty \frac{1}{\sqrt{\theta}} \exp(-\theta x) d\theta$$

$$= e^x \int_{\sqrt{2x}}^\infty \sqrt{2/x} \exp(-t^2/2) dt$$

$$= \sqrt{2/x}\ \Phi(-\sqrt{2x})/\phi(\sqrt{2x})$$

hence

$$(4.26) \qquad J_{-\frac{1}{2}}(x) = \sqrt{2/x}\ R(\sqrt{2x}),\ x \in \mathbb{R}^+$$

More generally, a relation will be established between $J_{-m-\frac{1}{2}}$ and the finite series, which FELLER 1950, p. 193, gave for Mills' ratio and which in fact traces back to LAPLACE 1812:

$$R(x) \approx \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{3*5}{x^7} + \frac{3*5*7}{x^9} - \dots$$

This series has the property that the partial sum of an even (odd) number of terms is always smaller (larger) than R(x) for all $x > 0$. The next lemma states this result more precisely.

Lemma 4.15. With the definitions

$$(4.27) \qquad t_k(x) := \frac{(-1)^k}{x^{2k+1}} \prod_{\ell=1}^{k} (2\ell-1),\ k \geq 1; \quad t_0(x) := \frac{1}{x}$$

$$(4.28) \qquad S_m(x) := \sum_{k=0}^{m} t_k,\ m \geq 0; \quad S_{-1}(x) := 0$$

Mills' ratio R(x) satisfies

$$(4.29) \qquad (-1)^m [R(x) - S_{m-1}(x)] > 0$$

for all $m \geq 0$ and all $x > 0$.

Proof. Inequality (4.29) immediately follows from the positivity of J_α and the next lemma. □

Lemma 4.16. With $y := \sqrt{2x}$, the equality

$$(4.30) \qquad J_{-m-\frac{1}{2}}(x) = 2[R(y) - S_{m-1}(y)]/[y^2 t_m(y)]$$

holds for all $m \geq 0$ and all $x > 0$.

Proof. Induction does the trick. For $m = 0$, (4.30) boils down to (4.26). Assume that (4.30) holds for a given $m \geq 0$ or, equivalently,

$$R(y) - S_{m-1}(y) = \frac{1}{2} y^2 t_m(y) J_{-m-\frac{1}{2}}(x)$$

Insertion of

$$\frac{1}{2} y^2 J_{-m-\frac{1}{2}}(y) = 1 - (m+\tfrac{1}{2}) J_{-m-3/2}(y)$$

which follows from Theorem 4.13 (i), gives

$$R(y) - S_{m-1}(y) = t_m(y) - (m+\tfrac{1}{2}) t_m(y) J_{-m-3/2}(y)$$

Since definition (4.27) implies

$$(4.31) \qquad t_{m+1}(x) = - \frac{2m+1}{x^2} t_m(x)$$

the equality

$$R(y) - S_m(y) = \frac{1}{2} y^2 t_{m+1}(y) J_{-(m+1)-\frac{1}{2}}(y)$$

immediately follows, completing the induction argument. □

Note that the proof of Lemma 4.15 is at least as simple as Feller's; compare also the proof in KINGMAN & TAYLOR 1966, p. 311. Moreover, relation (4.30) can be used to improve inequality (4.29), leading to sharper bounds for Mills' ratio. Theorems 4.17 and 4.19 show the way.

Theorem 4.17. Mills' ratio R(x) satisfies

$$(4.32) \qquad [R(x) - S_{m-1}(x)]/t_m(x) > \frac{x^2}{x^2+2m+1}$$

for all $m \geq 0$ and all $x > 0$, where t_m and S_m are defined by (4.27) and (4.28) respectively.

Proof. In view of Theorem 4.13 (ii), $J_{-m-\frac{1}{2}}$ exceeds $J_{-m-3/2}$ for all positive x. Lemma 4.16 then implies:

$$t_m(y)[R(y) - S_m(y)] > t_{m+1}(y)[R(y) - S_{m-1}(y)]$$

or, equivalently,

$$[t_m(y) - t_{m+1}(y)][R(y) - S_{m-1}(y)] > t_m^2(y)$$

which is identical with (4.32) by virtue of (4.31). □

Comparison of (4.32) and (4.29) immediately shows that Feller's inequality is improved indeed. The degree of improvement can be measured by comparing the difference between a consecutive upper and lower bound for R(x). Let therefore $M_{m-1}(x)$ denote the approximation for R(x), arising from viewing (4.32) as an equality:

$$(4.33) \qquad M_{m-1}(x) := S_{m-1}(x) + \frac{x^2}{x^2+2m+1} t_m(x)$$

For even (odd) m, $M_m(x)$ presents an upper (lower) bound for R(x). Now, the quantity

$$(4.34) \qquad V_m(x) := \frac{M_m(x)-M_{m-1}(x)}{S_m(x)-S_{m-1}(x)}$$

measures for $m \geq 0$ the improvement of the new approximation over Feller's. Direct calculation gives

$$(4.35) \qquad V_m(x) = \frac{2(2m+1)}{(x^2+2m+1)(x^2+2m+3)}$$

implying a very sizable improvement.

By use of (4.31), (4.33) may be written as

$$(4.36) \qquad M_m(x) = S_{m-1}(x) + \frac{x^2+2}{x^2+2m+3}\, t_m(x)$$

for $m \geq 0$, which expression simplifies the comparison with Feller's approximations. For small values of m, the approximations considered here are listed below.

$$S_{-1}(x) = 0 \qquad\qquad M_{-1}(x) = \frac{x}{x^2+1}$$

$$S_0(x) = \frac{1}{x} \qquad\qquad M_0(x) = \frac{1}{x} \cdot \frac{x^2+2}{x^2+3}$$

$$S_1(x) = \frac{1}{x} - \frac{1}{x^3} \qquad\qquad M_1(x) = \frac{1}{x} - \frac{1}{x^3} \cdot \frac{x^2+2}{x^2+5}$$

$$S_2(x) = \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} \qquad\qquad M_2(x) = \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} \cdot \frac{x^2+2}{x^2+7}$$

In Figure 4.18 Mills' ratio and its different approximations for $x > 0$ are shown. The lower bound $M_{-1}(x)$ was earlier derived by GORDON 1941.

Even further improvements are possible by exploiting the properties of the integrals J_α.

Theorem 4.19. Mills' ratio R(x) satisfies

$$(4.37) \qquad [R(x) - S_{m-1}(x)]/t_m(x) < \frac{x^4+(2m+1)x^2}{x^4+2(2m+1)x^2+(2m-1)(2m+1)}$$

for all $m \geq 1$ and all $x > 0$, where t_m and S_m are defined by (4.27) and (4.28) respectively.

Proof. Using Theorem 4.13 (iv) twice gives

$$J''_{-m-3/2}(x) = J_{-m-3/2}(x) - 2J_{-m-\frac{1}{2}}(x) + J_{-m+\frac{1}{2}}(x)$$

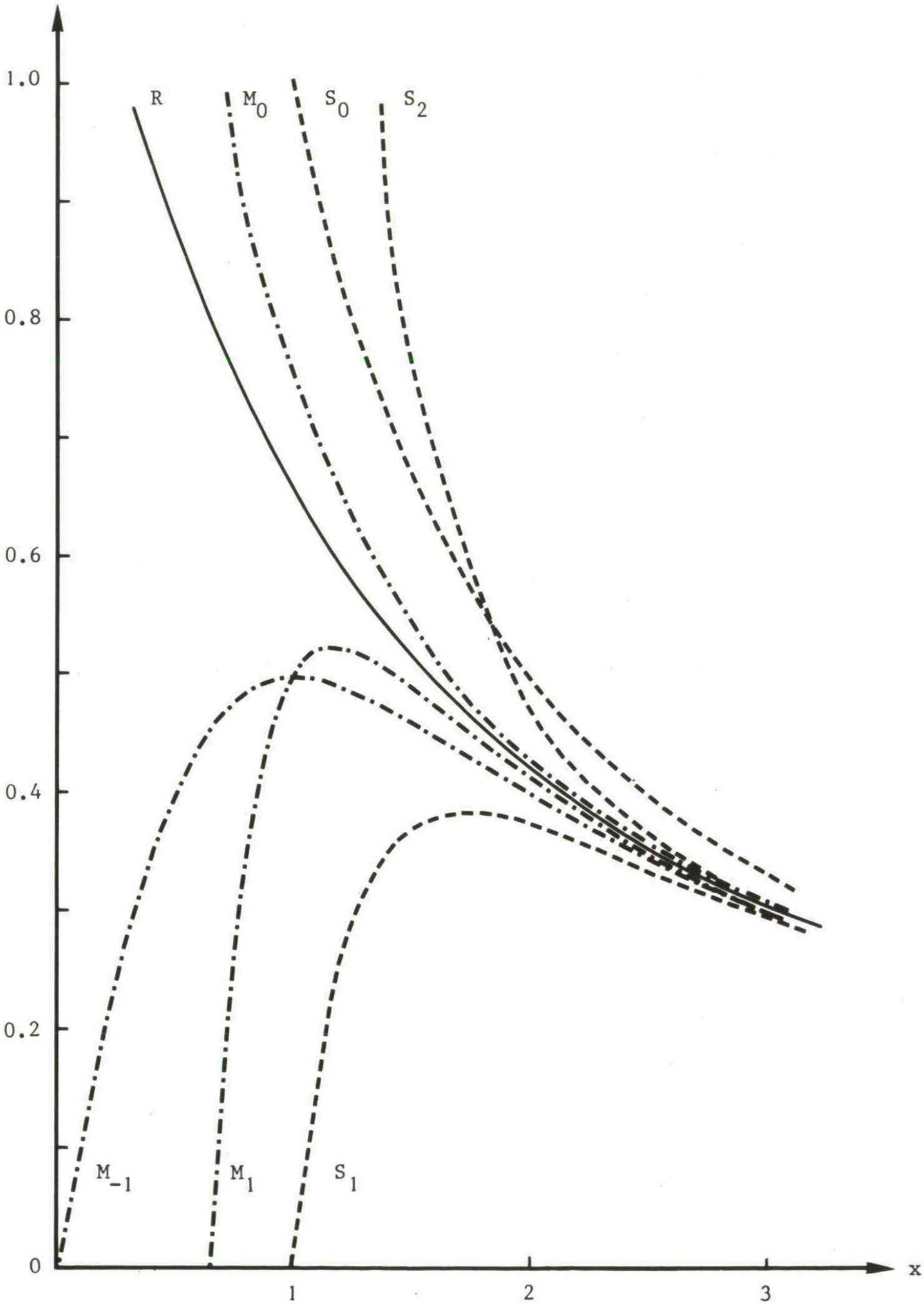


Figure 4.18. Different approximations to Mills' ratio

which is positive. Inserting (4.30) in the right hand side implies for $m \geqslant 1$ the positivity of

$$\frac{R(y)-S_{m-1}(y)-t_m(y)}{t_{m+1}(y)} - 2\,\frac{R(y)-S_{m-1}(y)}{t_m(y)} + \frac{R(y)-S_{m-1}(y)+t_{m-1}(y)}{t_{m-1}(y)}$$

which may be rewritten as

$$\frac{R(y)-S_{m-1}(y)}{t_m(y)}\left[\frac{t_m(y)}{t_{m+1}(y)} + \frac{t_m(y)}{t_{m-1}(y)} - 2\right] > \frac{t_m(y)}{t_{m+1}(y)} - 1$$

Application of (4.31) completes the proof. □

With the definition

$$(4.38) \qquad N_{m-2}(x) := S_{m-1}(x) + \frac{x^4+(2m+1)x^2}{x^4+2(2m+1)x^2+(2m-1)(2m+1)}\, t_m(x)$$

for $m \geqslant 1$, new bounds for $R(x)$ have been obtained. Rewriting (4.38) leads to

$$(4.39) \qquad N_m(x) = S_{m-1}(x) + \frac{x^4+(2m+9)x^2+8}{x^4+2(2m+5)x^2+(2m+3)(2m+5)}\, t_m(x)$$

for $m \geqslant 0$. The first approximations to $R(x)$ are:

$$N_{-1}(x) = \frac{x(x^2+5)}{x^4+6x^2+3}$$

$$N_0(x) = \frac{1}{x} \cdot \frac{x^4+9x^2+8}{x^4+10x^2+15}$$

$$N_1(x) = \frac{1}{x} - \frac{1}{x^3} \cdot \frac{x^4+11x^2+8}{x^4+14x^2+35}$$

$$N_2(x) = \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} \cdot \frac{x^4+13x^2+8}{x^4+18x^2+63}$$

The improvement with respect to Feller's approximations can be measured by

$$W_m(x) := \frac{N_m(x)-N_{m-1}(x)}{S_m(x)-S_{m-1}(x)} = \frac{8(2m+1)(2m+3)}{H_{2m+1}(x)H_{2m+3}(x)}$$

where

$$H_n(x) := x^4 + 2(n+2)x^2 + n(n+2)$$

Note that all new approximations considered here consist of a correction of the last term in Feller's approximation $S_m(x)$. It follows that for small values of x the bounds for R(x) are not very satisfactory. For $0 < x < 1$, the approximations given by BIRNBAUM 1942, SAMPFORD 1953, BOYD 1959 (or MOORS 1973) are far better.

PART 2

BINOMIAL DISTRIBUTION

5. TRUNCATED BINOMIAL PROBLEMS

5.1. Introduction and summary

In the remaining three chapters truncated estimation problems (Θ,L,X) will be discussed, where the observable random variable X has the binomial distribution $B(n,\theta)$. It is known beforehand that the unknown parameter θ belongs to an interval that is strictly smaller than the natural parameter space (0,1). In general, the loss function will be the quadratic $[h(\theta) - a]^2$, the exception being Section 7.4 where a weighted quadratic loss function will be considered. For easy reference, this kind of estimation problems receives a name of its own.

Definition 5.1. An estimation problem (Θ,L,X) with $\Theta := (Q,P) \subsetneq (0,1)$, $L(\theta,a) = [h(\theta) - a]^2$ and $X \in B(n,\theta)$ is called a truncated binomial problem; furthermore, it is called symmetric, if $Q + P = 1$. □

If follows at once that truncated binomial problems are quadratic in the sense of Definition 1.5. Hence, Corollary 1.6 is applicable, giving the following result.

Corollary 5.2. For truncated binomial problems class D of nonrandomized rules is essentially complete. □

Truncated binomial problems of course occur when in a given population the fraction θ with a certain characteristic has to be estimated on the base of a simple random sample, while external prior knowledge guarantees that θ cannot be lower than Q or higher than P. This situation is no exception at all. For example, consider the estimation of the probability θ of throwing heads with a given coin; it is the author's conviction that it is technically impossible to manufacture a coin

(i) that passes for fair at a superficial inspection, and

(ii) for which the probability of throwing heads is lower than 0.4 or higher than 0.6.

Some small-scale experiments with artificial coins support this view.

Other decision problem that can be rewritten as truncated binomial problems arise in the setting of randomized response. In that case the truncation of the parameter space is determined objectively: it follows logically from the nature of the decision problem, without any subjective consideration entering into the argument. By way of illustration, Warner's randomized response method was discussed in the Introduction.

Randomized response methodology therefore constitutes an interesting field of application of truncated estimation theory, in particular truncated binomial problems. For that reason a survey of randomized response methods and applications is presented in Section 5.2. Two basic randomized response techniques are discussed in some detail, namely Warner's original method and Simmons' unrelated question model; both techniques will return to the scene in Section 5.3. Further, attention focuses on more recent developments in the field, since the survey article by HORVITZ et al. 1976 adequately summarizes the older contributions.

In Section 5.3 the general Theorem 2.3 concerning the existence of an unbiased estimator is applied to the present situation. The remainder of Section 5.3 concentrates on admissibility: Theorem 3.9 about the inadmissiblity of boundary rules in linvariant problems is applied to both Warner's and Simmons' randomized response method.

Bayes rules are treated in Section 5.4, in particular with respect to the conjugated class of prior distributions, consisting of truncated beta distributions. The limiting cases of this class are discussed in some detail as well.

An introduction to the derivation of minimax rules is presented in the final Section 5.5. It is shown in Section 5.4 that the Bayes risk only depends on a finite number of moments of the prior distribution. Since finding a minimax rule is equivalent to deriving a least favorable prior distribution, it is necessary to know what values the vector consisting of the first moments can take for distributions on Θ. This leads to the celebrated moment problem, to be discussed in Chapter 6. The actual calculation of minimax rules is reported in the final Chapter 7.

Truncated binomial problems have been studied before, of course. BLUM & ROSENBLATT 1967 considered T-minimaxity in the parameter space $[0,P]$ and SCHAFER 1976 offered some alternatives to the usual minimax estimator. RAFSKY 1976 noted that for finite populations the parameter space is in fact discrete; this led him to an estimator dominating the sample proportion. Finally, SKIBINSKY & COTE 1963 can be mentioned; they considered the case where the prior knowledge implies that θ is outside the interval $[1-P,P]$ with a given (small) probability only.

5.2. Randomized response

To reduce reluctance of interviewees to cooperate in surveys about sensitive personal matters, WARNER 1965 devised an interview method that better protected the respondent's privacy. The respondent draws at random one out of the two statements

'I possess property A'

'I do not possess property A'

and answers 'correct' or 'false'. The interviewer is not allowed to know which statement was drawn, so that the respondent's true situation is not revealed. If A is a sensitive or incriminating property (having had an abortion is the example often used), it seems plausible that refusals to answer or untruthful replies will occur less frequently. This interview technique will be called Warner's method of randomized response.

The probability with which the first statement (having property A) is drawn, will be denoted by P, a constant between 0 and 1 to be chosen by the statistician. The probability of drawing the second statement then is $1-P$. Without loss of generality $\frac{1}{2} < P \leq 1$ may be assumed. (For $P = \frac{1}{2}$ the estimation problem to follow is not identifiable.) Note that for $P = 1$ in fact a nonrandomized statement is presented. Suppose that a random sample of size n is drawn with replacement from the population under investigation and let θ denote the probability of receiving the answer 'correct'. Then the observable random variable X is the number of sample persons who answered 'correct'; furthermore, X has the binomial distribution $B(n,\theta)$. Assume that under these circumstances all respondents reply truthfully; then θ satisfies

(5.1) $\theta = (1-P) + (2P-1)\pi_A$

where π_A is the parameter of interest: the population fraction with property A. From $0 \leq \pi_A \leq 1$ the double inequality

(5.2) $1 - P \leq \theta \leq P$

follows at once, so that for $P \neq 1$ the parameter space is truncated indeed. (Compare also the Introduction.) Attention will be concentrated on the estimation of π_A; since quadratic loss will be used, this is equivalent to estimating $\pi_A - \frac{1}{2}$. Later on, the latter choice will appear to simplify the analysis somewhat, hence the estimand is chosen to be

(5.3) $h(\theta) = \pi_A - \frac{1}{2} = (\theta-\frac{1}{2})/(2P-1)$

Note that the estimation problem indeed proves to be a (symmetric) truncated binomial problem.

A significant modification and generalization of Warner's method was suggested by Simmons; see for a detailed description HORVITZ et al. 1967 and GREENBERG et al. 1969. His idea was to replace the second of Warner's two statements by

'I possess property Y'

where Y is some other characteristic, which causes no embarrassment and has no condemning quality whatsoever. It is thought that the respondent's confidence in the protection of his privacy will be increased by providing him/her the opportunity to reply to a completely harmless question, which bears no relation to the sensitive property A. The method is known as the unrelated question randomized response technique or Simmons' method. The unrelated property Y can be chosen by the statistician. Whether the population fraction π_Y showing property Y is known to the statistician or not makes an important difference for the estimation problem; therefore, two cases will be distinguished:

(i) π_Y is known to the statistician;

(ii) π_Y is unknown.

The two cases will be briefly discussed now.

(i) If the statistician knows the value of π_Y beforehand, the problem very much resembles Warner's method. Again, the observable number X of answers 'correct' has the distribution $B(n,\theta)$, where now, however,

$$(5.4) \qquad \theta = P\pi_A + (1-P)\pi_Y$$

Hence, $\Theta = [(1-P)\pi_Y, P + (1-P)\pi_Y]$ and for $P \neq 1$ the choice

$$(5.5) \qquad h(\theta) = \pi_A - \tfrac{1}{2} = [\theta - P/2 - (1-P)\pi_Y]/P$$

gives rise to a truncated binomial problem again. Note that it is symmetric only for $\pi_Y = \frac{1}{2}$.

(ii) If π_Y is unknown, it is a nuisance parameter which has to be estimated as well. The most obvious method is to take two independent random samples from the population and use different fractions of the two statements in each sample. If the two samples and their characteristics are indicated by suffices 1 and 2 respectively, the probabilities of obtaining the answer 'correct' are

$$\theta_i = P_i\pi_A + (1-P_i)\pi_Y$$

for $i = 1,2$. Hence the unknown parameter $\theta := (\theta_1,\theta_2)$ satisfies

$$(5.6) \qquad \theta = \begin{pmatrix} P_1 & 1-P_1 \\ P_2 & 1-P_2 \end{pmatrix} \begin{pmatrix} \pi_A \\ \pi_Y \end{pmatrix}$$

Figure 5.3 shows the parameter space Θ for typical values of P_1 and P_2; again, a truncated decision problem turns up, where vector $(\pi_A-\frac{1}{2}, \pi_Y-\frac{1}{2})$ can be taken as estimand:

$$(5.7) \qquad h(\theta) = \frac{1}{P_1-P_2} \begin{pmatrix} 1-P_2 & -(1-P_1) \\ -P_2 & P_1 \end{pmatrix} \begin{pmatrix} \theta_1-\frac{1}{2} \\ \theta_2-\frac{1}{2} \end{pmatrix}$$

Note that it is necessary to take P_2 unequal to P_1; otherwise the estimation problem is not identifiable and the estimation procedure breaks down.

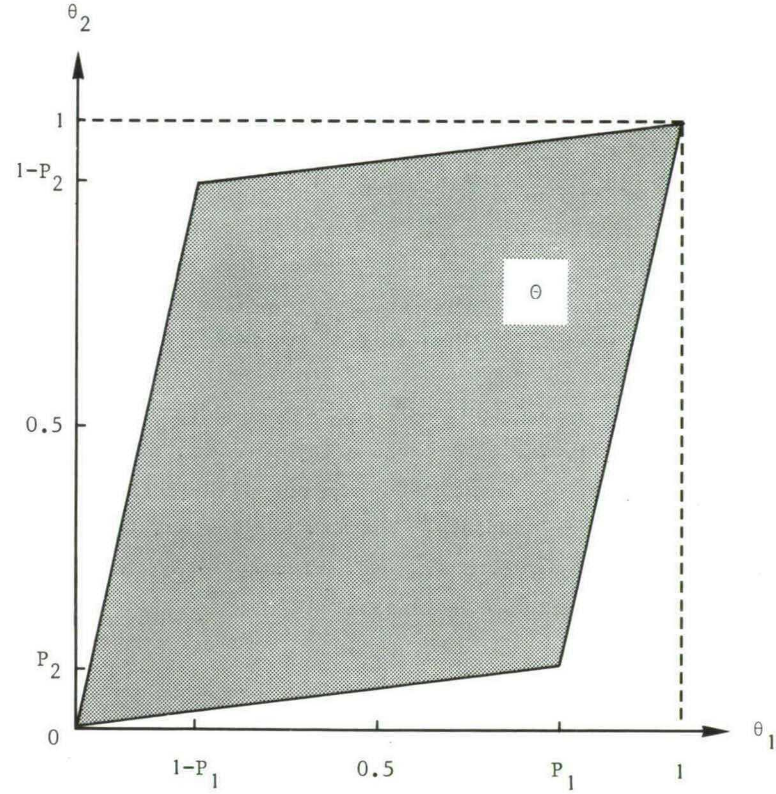


Figure 5.3. (Truncated) parameter space for Simmons' method of randomized response with π_Y unknown

Simmons' methods leave the statistician some choices: he has to choose the unrelated property Y (with the important question whether π_Y is known or not included) and he must pick the design constants P_1 and P_2 as well as the allocation (if π_Y is unknown). GREENBERG et al. 1969 discussed these matters; improvements and additions were given by MOORS 1971 and LANKE 1975.

During the congress of the International Statistical Institute in Warsaw 1975, a seperate session was devoted to randomized response methodology. The proceedings of this meeting were published in the International Statistical Review 44/2 of August 1976. The interested reader is referred to the papers in this issue, particularly to the exhaustive survey paper by HORVITZ et al. 1976. For Dutch readers the survey by VERDOOREN 1976 can be recommended as well.

Of more recent developments, the following are considered the most interesting. Independently and almost simultaneously, several authors pointed out that while comparing different randomized response methods, the degree of privacy protection is as important as the variance of the resulting estimates. A balance must be found between two conflicting interests: on the one hand the statistician who strives for efficient estimates, on the other hand the respondent who wants to protect his/her privacy as much as possible. This approach was taken by LANKE 1976, LEYSIEFFER & WARNER 1976 and ANDERSON 1977 for example. Compare also FLIGNER et al. 1977.

Practical applications of randomized response methods in large scale surveys are not abundant, although extremely interesting. The following papers also make empirical comparisons between several interview methods: KROTKI & FOX 1974 or KROTKI & McDANIEL 1975 (on abortions), GOODSTADT & GRUSON 1975 (drug use), LOCANDER et al. 1976 (bankruptcy, drunken driving, a.o.) and SHIMIZU & BONHAM 1978 (abortions). It was unanimously reported that the use of randomized response methods resulted in fewer refusals to cooperate and less underreporting of sensitive properties. A striking illustration is Table 5.4, taken from KROTKI & McDANIEL 1975, who also report a number of 4,040 legitimate therapeutic abortions.

Table 5.4. Estimated induced abortions (therapeutic and illegal), Alberta 1973.

Randomized response technique	12,320
Questionnaire	3,060
Interview	1,150

Note that the estimates obtained by the two conventional data gathering methods are even below the number of legitimate abortions!

Finally, with some imagination the quite different problem discussed by DOWNS et al. 1978 can be seen as an application of randomized response.

Randomized response models can be used to estimate not only population fractions (means of dichotomous variables), but also means of quantitative variables. POLLOCK & BEK 1976 compared the three methods, suggested previously by GREENBERG et al. 1969, WARNER 1971 and POOLE 1974, with respect to their efficiency. ALBERS 1982 took into consideration the privacy aspect as well and proposed another, more refined method. His model combines the suggestions of Greenberg en Warner mentioned above and seems to satisfy all reasonable desiderata.

If 'unrelated property' Y in Simmons' method is sensitive as well, it is possible to estimate the incidence of two sensitive properties simultaneously. Moreover, the association between the two characteristics can be estimated; even tests of dependence are possible. DRANE 1975 and CLICKNER & IGLEWICZ 1976 were the first to follow this line of thought; again, many modifications present themselves. TAMHANE 1981 thoroughly discussed this approach and compared the competing techniques. Compare also BOURKE 1982.

For the methods named after Warner and Simmons, the simplest estimates for π_A (and π_Y) are obtained by plugging into equation (5.1), (5.4) or (5.6) the observed fractions of answers 'correct', and solving the unknowns. That these estimates must be truncated at the boundaries of the parameter space to produce maximum likelihood estimators, was noted by DEVORE 1977 and FLIGNER et al. 1977. Guided by a simple numerical example, RAGHAVARAO 1978 concluded that even the MLE is inadmissible for Warner's method and suggested an estimate based on a logistic transformation. Bayes' estimators were discussed by WINKLER & FRANKLIN 1979, using as prior distributions the truncated beta distributions, introduced here in Definition 2.9.

BELLHOUSE 1982 further generalized the linear randomized response model, presented by WARNER 1971, and derived some general optimality results.

5.3. Admissibility

Without further ado, Theorem 2.3 can be applied to the estimation problems introduced up to now in this chapter. For truncated binomial problems it follows at once that no unbiased estimators exist, since an unbiased estimator exists for θ, $\theta \in (0,1)$. Hence, the same holds for both Warner's method of randomized response and Simmons' method with known π_Y (except in the trivial case $P = 1$). Simmons' method with unknown π_Y is not equivalent to a truncated binomial problem; nevertheless, Theorem 2.3 again implies that an unbiased estimator only exists in the (nonrandomized) case $P_1 = P_2 = 1$. These results are summarized in the next theorem.

Theorem 5.5. For Warner's nor for Simmons' randomized response models an unbiased estimator exists, except in the (effectively nonrandomized) cases $P = 1$ or $P_1 = P_2 = 1$. □

Inadmissibility of boundary rules was previously discussed for the special case of a truncated binomial problem with one observation and with h the identity: Example 2.4 showed that estimators taking values exactly on the boundary of any two-sidedly truncated parameter space necessarily are inadmissible; if the parameter space includes the value $\frac{1}{2}$, then even rules with values close to the boundary were seen to be inadmissible in Example 2.6. This last result will be generalized now by means of Theorem 3.9. Since the randomized response models present practical applications, attention is focused on them.

Theorem 5.6. Consider estimation of a population fraction π_A with quadratic loss by means of Warner's method of randomized response. Let d be a linvariant estimator, taking values outside the interval with endpoints $\frac{1}{2}$ and $\frac{1}{2} + \frac{1}{2}\tanh[(x-n/2)\log\{P/(1-P)\}]$ for some $x = 0,1,\ldots,n$. Then d is inadmissible.

Proof. Consider the estimation problem with $L(\theta,a) = [h(\theta) - a]^2$ where h is defined by (5.3). This problem is linvariant under the group of two elements with $g(x) = n-x$, $\bar{g}(\theta) = 1-\theta$ and $\tilde{g}(a) = -a$.
Since $X \in B(n,\theta)$, (3.5) results in

$$h_x(\theta) = h(\theta)\,\frac{f(x|\theta)-f(x|\bar{g}(\theta))}{f(x|\theta)+f(x|\bar{g}(\theta))} = h(\theta)\,\frac{\theta^{2x-n}-(1-\theta)^{2x-n}}{\theta^{2x-n}+(1-\theta)^{2x-n}}$$

or

(5.8) $\quad h_x(\theta) = h(\theta) \tanh[(x-n/2) \log\{\theta/(1-\theta)\}]$

Extremes are obtained for $\theta = \frac{1}{2}$ and $\theta = P$. For $x > n/2$, these extremes are the minimum and the maximum, respectively; for $x < n/2$ the situation is reversed. So, $A_x = h_x(\Theta)$ is the interval with endpoints 0 and $\frac{1}{2} \tanh[(x-n/2) \log\{P/(1-P)\}]$; by virtue of Theorem 3.9 linvariant estimators taking values outside this interval are inadmissible. Finally, the estimation problem for π_A comes down to a translation over $\frac{1}{2}$ and is equivalent otherwise. □

The inadmissible set (cf. Corollary 3.11) for this problem is the shaded area in Figure 5.7. It follows at once that the usual estimator for π_A, based upon the sample fraction of answers 'correct', is inadmissible; what is more: the same holds for the MLE.

The curve in Figure 5.7 can be interpreted as the graph of an estimator for π_A. In the next section it will be shown that this estimator is precisely the Bayes estimator corresponding with a limiting member of the conjugated family.

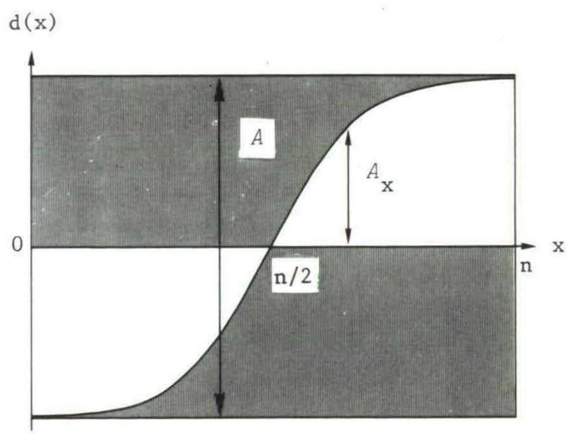


Figure 5.7. Inadmissible set (shaded area) for Warner's method of randomized response; case n = 14 and P = 0.6

The probability that the inadmissible set V_X contains π_A can be found as follows: for $\theta > \frac{1}{2}$

$$P_\theta\{\theta \in V_X\} = P_\theta\{\theta > \tfrac{1}{2} + \tfrac{1}{2}\tanh[(X-n/2)\log\{P/(1-P)\}]$$

$$= P_\theta\{X < n/2 + \tfrac{1}{2}\log[\theta/(1-\theta)]/\log[P/(1-P)]\}$$

$$= P_\theta\{X \leqslant \mathrm{ent}(n/2)\}$$

where ent denotes the entier function: ent(x) is the largest integer not exceeding x.

Finally, the probability that the MLE for π_A is in V_X can be approximated by $P_\theta\{X \leqslant n(1-P)\} + P_\theta\{X \geqslant nP\}$. For n not too small this is about equal to

$$\Phi\left(\frac{1-P-\theta}{\sqrt{\theta(1-\theta)}}\sqrt{n}\right) + \Phi\left(\frac{\theta-P}{\sqrt{\theta(1-\theta)}}\sqrt{n}\right)$$

which is small if θ is not close to the endpoints of Θ.

Next, consider Simmons' method of randomized response, where the population fraction π_Y with unrelated characteristic Y is known in advance. First take the case that π_Y equals $\tfrac{1}{2}$. Following the results in Section 5.2, the parameter space is then interval $[(1-P)/2,\ (1+P)/2]$, which is symmetric around $\tfrac{1}{2}$. Just as in Warner's model, the estimation problem is linvariant; h_X is again given by (5.8). Estimates for π_A must now be confined to the interval with endpoints $\tfrac{1}{2}$ and $\tfrac{1}{2} + \tfrac{1}{2}\tanh[(x-n/2)\log\{(1+P)/(1-P)\}]$. Note that the inadmissible set is a strict subset of the inadmissible set derived in Theorem 5.6 for Warner's method, reflecting the fact that now the parameter space is a wider interval. The result is formalized into a theorem.

Theorem 5.8. Consider estimation of a population fraction π_A with quadratic loss by means of Simmons' method of randomized response where π_Y is known to be $\tfrac{1}{2}$. Let d be a linvariant estimator, taking values outside the interval with endpoints $\tfrac{1}{2}$ and $\tfrac{1}{2} + \tfrac{1}{2}\tanh[(x-n/2)\log\{(1+P)/(1-P)\}]$ for some $x = 0,1,\ldots,n$. Then d is inadmissible. □

If in Simmons' method π_Y is known, but unequal to $\tfrac{1}{2}$, the estimation problem no longer is invariant. Note, however, that for

$$\frac{1-2P}{2(1-P)} < \pi_Y < \frac{1}{2(1-P)}$$

the parameter space contains $\frac{1}{2}$ as interior point and the problem is pseudo-invariant according to Definition 3.25.

Finally, the variant of Simmons' randomized response model will be discussed, where π_Y is unknown beforehand. This situation was briefly discussed in Section 5.2, where two independent random samples were used. The problem was parameterized by (5.6); Figure 5.3 showed the truncated parameter space. Application of Theorem 3.9 now leads to the following result.

Theorem 5.9. Consider estimation of a population fraction π_A with quadratic loss by means of Simmons' method of randomized response with π_Y unknown. Let d be a linvariant estimator for (π_A, π_Y), which for some observations x takes values outside the closed convex hull of the space bounded by the two curves

$$(5.9) \qquad \begin{pmatrix} \frac{1}{2} \\ \\ \frac{1}{2} \end{pmatrix} + k_1(\theta_1,x) \begin{pmatrix} \theta_1/P_1-\frac{1}{2} \\ \\ -\frac{1}{2} \end{pmatrix}, \quad 0 < \theta_1 \leq P_1$$

$$(5.10) \qquad \begin{pmatrix} \frac{1}{2} \\ \\ \frac{1}{2} \end{pmatrix} + k_2(\theta_1,x) \begin{pmatrix} -\frac{1}{2} \\ \\ \theta_1/(1-P_1)-\frac{1}{2} \end{pmatrix}, \quad 0 < \theta_1 \leq 1-P_1$$

where

$$k_1(\theta_1,x) := \tanh[(x_1-n_1/2)\log\frac{\theta_1}{1-\theta_1} + (x_2-n_2/2)\log\frac{P_2\theta_1}{P_1-P_2\theta_1}]$$

$$k_2(\theta_1,x) := \tanh[(x_1-n_1/2)\log\frac{\theta_1}{1-\theta_1} + (x_2-n_2/2)\log\frac{(1-P_2)\theta_1}{1-P_1-(1-P_2)\theta_1}]$$

Then d is inadmissible.

Proof. Let X_i denote the number of answers 'correct' in sample i, so that $X_i \in B(n_i,\theta_i)$ for i = 1,2. The independence gives

$$f(x|\theta) = \prod_{i=1}^{2} \binom{n_i}{x_i} \theta_i^{x_i} (1-\theta_i)^{n_i - x_i}$$

as probability mass function for (X_1,X_2) with observed value $x := (x_1, x_2)$. It is obvious that the class of distributions $\{f(x|\theta) : \theta \in \Theta\}$ is invariant under the group of two elements determined by

$$g(x) = \begin{pmatrix} n_1 - x_1 \\ n_2 - x_2 \end{pmatrix}$$

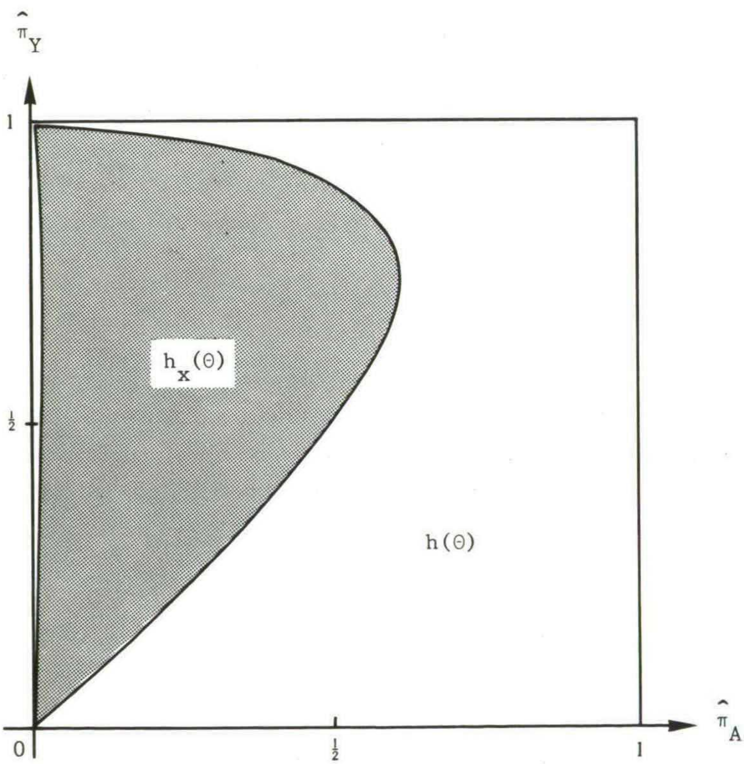


Figure 5.10. Restricted action space (shaded area) for Simmons' randomized response method with π_Y unknown and $P_1 = 0.8$, $P_2 = 0.1$; case $x_1-n_1/2 = -2$ and $x_2-n_2/2 = 1$

The problem of estimating (5.7) then is linvariant with

$$\bar{g}(\theta) = \begin{pmatrix} 1-\theta_1 \\ \\ 1-\theta_2 \end{pmatrix}, \quad \tilde{g}(a) = -a$$

and (3.5) leads to

$$h_x(\theta) = \tanh[(x_1-n_1/2)\log\frac{\theta_1}{1-\theta_1} + (x_2-n_2/2)\log\frac{\theta_2}{1-\theta_2}]h(\theta)$$

To find the space $h_x(\Theta)$ it suffices to consider the h_x-image of the boundaries of Θ. In combination with the translation needed to estimate (π_A,π_Y) itself, this gives the result stated in the theorem. □

Figure 5.10 shows a typical example of the restricted action space and the inadmissible set. Note that again the MLE is inadmissible in this variation of Simmons' randomized response method.

5.4 Bayes rules

In this section Bayes rules for the truncated binomial problem will be discussed. To start with, h is taken to be the identity, while an arbitrary prior distribution $\tau \in \Theta^*$ is considered. The following notations will be used for $j = 0,1,2,\ldots$ and $i = 0,1,\ldots,n$:

(5.11) $\mu_j := E_\tau(\theta^j)$

(5.12) $m_{ij} := E_\tau[\theta^{i+j}(1-\theta)^{n-i}]$

Note that the dependence on τ (and n) has been suppressed in these notations. Relation

(5.13) $\mu_j = \sum_{i=0}^{n} \binom{n}{i} m_{ij}\ , \ j = 0,1,2,\ldots$

is obvious.

Theorem 5.11. For the truncated binomial problem with h the identity, the nonrandomized Bayes rule d_τ with respect to some $\tau \in \Theta^*$ is given by

$$(5.14) \qquad d_\tau(i) = m_{i1}/m_{i0}\ , \ i = 0,1,\dots,n$$

This rule is admissible; its (minimum) Bayes risk equals

$$(5.15) \qquad r(\tau,d_\tau) = \mu_2 - \sum_{i=0}^{n} \binom{n}{i} m_{i1}^2/m_{i0}$$

Proof. Any nonrandomized rule d can be written as a (n+1)-tuple (d(0), d(1),...,d(n)) with $d(i) \in \Theta$ for all $i = 0,1,\dots,n$. The risk function is

$$R(\theta,d) = \sum_{i=0}^{n} \binom{n}{i} \theta^i(1-\theta)^{n-i}[\theta-d(i)]^2$$

Hence the Bayes risk of d with respect to some $\tau \in \Theta^*$ equals

$$r(\tau,d) = \sum_{i=0}^{n} \binom{n}{i}[m_{i2} - 2m_{i1}d(i) + m_{i0}d^2(i)]$$

$$= \sum_{i=0}^{n} \binom{n}{i}[m_{i2} - \frac{m_{i1}^2}{m_{i0}} + m_{i0}\{d(i) - \frac{m_{i1}}{m_{i0}}\}^2]$$

which is minimized by (5.14). Note that indeed $d_\tau(i) \in \Theta$ for $i = 0,1,\dots,n$; admissibility follows by Theorem 1.8. Finally, by virtue of (5.13) for $j = 2$, the minimum Bayes risk may be written as in (5.15). Note that (5.14) immediately follows from (4.13) as well. □

More explicit results can be obtained by considering the symmetric truncated binomial problem and, besides, the conjugated family of prior distributions. This situation was discussed in Example 2.10; the conjugated family was shown to consist of truncated beta distributions $B(\alpha,\beta;P)$.

Lemma 5.12 lists some straightforward properties of the central function $C(\alpha,\beta)$ defined in (2.10), while Lemma 5.13 presents two limiting results.

Lemma 5.12. For the function $C(\alpha,\beta)$ in (2.10) the following recursive relations hold

(i) $C(\alpha,\beta) = C(\beta,\alpha)$

(ii) $C(\alpha,\beta+1) + C(\alpha+1,\beta) = C(\alpha,\beta)$

(iii) $\alpha C(\alpha,\beta+1) - \beta C(\alpha+1,\beta) = P^{\alpha}(1-P)^{\beta} - P^{\beta}(1-P)^{\alpha}$

(iv) $(\alpha+\beta)C(\alpha+1,\beta) = \alpha C(\alpha,\beta) - [P(1-P)]^{\beta}[P^{\alpha-\beta} - (1-P)^{\alpha-\beta}]$

(v) $C(\alpha+1,\alpha) = C(\alpha,\alpha)/2$

(vi) $C(\alpha+2,\alpha) = C(\alpha+1,\alpha) - C(\alpha+1,\alpha+1)$

for all $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ and $\frac{1}{2} < P < 1$.

Proof. Properties (i) and (ii) are straightforward, while (iii) is obtained by partial integration. Combining (ii) and (iii) leads to (iv), while (v) and (vi) are special cases of (ii). □

Lemma 5.13. The following limits hold:

$$(5.16) \qquad \lim_{\alpha\to\infty} \frac{C(\alpha+x,\alpha+y)}{C(\alpha,\alpha)} = 2^{-x-y}$$

$$(5.17) \qquad \lim_{\alpha\to-\infty} \frac{C(\alpha+x,\alpha+y)}{C(\alpha,\alpha)} = [P^{x}(1-P)^{y} + P^{y}(1-P)^{x}]/2$$

for $x = 0,1,2,\ldots$ and $y = 0,1,2,\ldots$.

Proof. Only (5.17) will be proved, as the proof of (5.16) is similar, but simpler. Without loss of generality $x \geq y$ may be assumed; induction with respect to x will be used. For $x = 0$ (hence $y = 0$) (5.17) is correct. Suppose it holds for all $x \leq k$, where $k = 0,1,2,\ldots$. Property (iv) of Lemma 5.12 implies

$$\frac{C(\alpha+k+1,\alpha+y)}{C(\alpha,\alpha)} = \frac{\alpha+k}{2\alpha+k+y}\frac{C(\alpha+k,\alpha+y)}{C(\alpha,\alpha)} - \frac{P^k(1-P)^y-P^y(1-P)^k}{(2\alpha+k+y)/\alpha}\frac{[P(1-P]^\alpha}{\alpha C(\alpha,\alpha)}$$

For all negative α, partial integration gives

$$\frac{\alpha C(\alpha,\alpha)}{[P(1-P)]^\alpha} = \int_{1-P}^{P} \alpha\left[\frac{\theta(1-\theta)}{P(1-P)}\right]^{\alpha-1}\frac{d\theta}{P(1-P)} = \int_{1-P}^{P}\frac{1}{1-2\theta}\,d\left[\frac{\theta(1-\theta)}{P(1-P)}\right]^\alpha$$

$$= \left[\frac{1}{1-2\theta}\left\{\frac{\theta(1-\theta)}{P(1-P)}\right\}^\alpha\right]_{1-P}^{P} - \int_{1-P}^{P}\left[\frac{\theta(1-\theta)}{P(1-P)}\right]^\alpha d\,\frac{1}{1-2\theta}$$

$$= -\frac{2}{2P-1} - 2\int_{1-P}^{P}(1-2\theta)^{-2}\left[\frac{\theta(1-\theta)}{P(1-P)}\right]^\alpha d\theta$$

Since $\theta(1-\theta)/\{P(1-P)\} > 1$ for $1-P < \theta < P$, the last integrand approaches 0 for $\alpha \to -\infty$ except in the isolated points $1-P$, $\frac{1}{2}$ and P. So the integral tends to 0 for $\alpha \to -\infty$, which implies

$$\lim_{\alpha\to-\infty}\frac{C(\alpha+k+1,\alpha+y)}{C(\alpha,\alpha)} = \frac{1}{2}\lim_{\alpha\to-\infty}\frac{C(\alpha+k,\alpha+y)}{C(\alpha,\alpha)} + [P^k(1-P)^y-P^y(1-P^k]\,\frac{2P-1}{4}$$

The lemma follows from the induction assumption (for $y = k+1$ first apply property (v) of Lemma 5.12). □

In the next theorem linvariant Bayes rules are derived for the case that h is the identity. Class $\{B(\alpha,\alpha;P) : \alpha \in \mathbb{R}\}$ of symmetric prior distributions is extended to include the two limiting cases $\alpha \to -\infty$ and $\alpha \to \infty$. It is plausible to define $B(-\infty,-\infty;P)$ as the two-point distribution, giving point mass $\frac{1}{2}$ to P and $1-P$ each, while $B(\infty,\infty;P)$ is the degenerate distribution with all mass concentrated in $\frac{1}{2}$. Compare the notation in Example 2.12.

Theorem 5.14. For a symmetric truncated binomial problem with h the identity, rules d_α defined for $x = 0,1,\ldots,n$ by

$$(5.18)\qquad d_\alpha(x) := \frac{C(\alpha+x+1,\alpha+n-x)}{C(\alpha+x,\alpha+n-x)}$$

are admissible, linvariant and Bayes with respect to $B(\alpha,\alpha;P)$ for $-\infty \leqslant \alpha \leqslant \infty$.

Proof. According to (5.14) Bayes rule d with respect to $B(\alpha,\beta;P)$ is given by

$$d(x) = \frac{C(\alpha+x+1,\beta+n-x)}{C(\alpha+x,\beta+n-x)}$$

Then, properties (i) and (ii) of Lemma 5.12 imply

$$1-d(n-x) = \frac{C(\beta+x+1,\alpha+n-x)}{C(\beta+x,\alpha+n-x)}$$

so that invariance is achieved for $\alpha = \beta$. Admissibility follows from Theorem 5.11. □

The limiting cases $d_{-\infty}$ and d_{∞} are found with Lemma 5.13:

$$(5.19) \qquad d_{-\infty}(x) = \tfrac{1}{2} + (P-\tfrac{1}{2}) \tanh[(x-n/2) \log \frac{P}{1-P}]$$

$$(5.20) \qquad d_{\infty}(x) = \tfrac{1}{2}$$

It is simple to apply these results to the randomized response methods of both Warner and Simmons (with $\pi_Y = \frac{1}{2}$), since the only difference arises from the fact that h no longer is the identity, but rather a linear function of θ. This difference is inessential, however: if d is Bayes for θ, then $ad+b$ is Bayes for $a\theta+b$ (a and $b \in \mathbb{R}$), as is immediately clear from (2.7). It is interesting to apply the appropriate transformation to the estimators (5.19) and (5.20) in the case of Warner's method: it can easily be seen that these two estimators correspond to the boundaries of the inadmissible set in Figure 5.7.

A generalization of the Bayes rules (5.18) was briefly discussed in MOORS 1977: prior distributions were considered that are a weighted mean of $B(\alpha,\alpha;P)$ and $B(\beta,\beta;P)$ with α and $\beta \in \mathbb{R}$. Truncated beta distributions in the setting of Warner's randomized response model were also discussed by WINKLER & FRANKLIN 1979.

5.5. Introduction to minimax rules

As was illustrated in Section 2.4, minimax rules for truncated estimation problems are much harder to find in general than in the classical, non-truncated case. For that reason, from now on attention will be focused on one specific problem: in the remainder of the book minimax rules will be found for the symmetric truncated binomial problem of Definition 5.1. Note that this problem was covered in Example 2.12 for the special case $n = 1$, while the classical, nontruncated case was treated in Example 1.14. This nontruncated problem is the limiting case $(Q = 0,\ P = 1)$ of the truncated problems in Definition 5.1. Hence, it is reasonable to expect that the minimax estimator

$$(1.12) \qquad d_m(x) = \frac{x+\sqrt{n}/2}{n+\sqrt{n}}\ ,\ x = 0,1,\ldots,n$$

in Example 1.14 will be minimax in the (symmetric) truncated situation as well, at least for P close to 1. The next theorem states sufficient conditions.

Theorem 5.15. Rule d_m defined by (1.12) is minimax for the symmetric truncated binomial problem with h the identity and with

$$(5.21) \qquad P > (n+\sqrt{n}/2)/(n+\sqrt{n})$$

if a probability distribution on $[1-P,P]$ exists with moments μ_i satisfying $\mu_0 = 1$ and

$$(5.22) \qquad \mu_{i+1} = \frac{i+\sqrt{n}/2}{i+\sqrt{n}}\,\mu_i\ ,\ i = 0,1,\ldots,n$$

Proof. Condition (5.21) is obvious from the necessity $d_m(n) < P$. Since minimax rule d_m is Bayes as well, Theorem 5.11 implies the existence of a distribution on $[1-P,P]$ satisfying

$$(5.23) \qquad m_{i1}/m_{i0} = (i+\sqrt{n}/2)/(n+\sqrt{n})$$

for $i = 0,1,\ldots,n$. The equivalence of (5.22) and (5.23), which is the subject of the next lemma, completes the proof. □

Lemma 5.16. The three following systems of equations are equivalent

$$\text{(i)}\quad \frac{E[\theta^{i+1}(1-\theta)^{n-i}]}{E[\theta^{i}(1-\theta)^{n-i}]} = \frac{i+\sqrt{n}/2}{n+\sqrt{n}}\ , \quad i = 0,1,\ldots,n$$

$$\text{(ii)}\quad \frac{E[\theta^{i+1}(1-\theta)^{j}]}{E[\theta^{i}(1-\theta)^{j}]} = \frac{i+\sqrt{n}/2}{i+j+\sqrt{n}}\ , \quad i = 0,1,\ldots,n;\ j = 0,1,\ldots,n-i$$

$$\text{(iii)}\quad \frac{E(\theta^{i+1})}{E(\theta^{i})} = \frac{i+\sqrt{n}/2}{i+\sqrt{n}}\ , \quad i = 0,1,\ldots,n$$

Proof. The lemma is proved following the scheme (iii) ⇒ (ii) ⇒ (i) ⇒ (iii).

(iii) ⇒ (ii). Denote the left-hand side of (ii) by $\phi(i,j)$. Relation

$$\phi(i,j+1) = \frac{1-\phi(i+1,j)}{1/\phi(i,j)-1} \quad i = 0,1,\ldots,n-j-1$$

is easy to establish, so the correctness of (ii) for some $j (0 \leq j \leq n)$ and all $i \leq n-j$ implies the correctness for $j+1$ and all $i \leq n-j-1$. Since (ii) and (iii) are identical for $j = 0$, repeated application gives the desired result.

(ii) ⇒ (i). Follows from substitution $j = n-i$.

(i) ⇒ (iii). The proof is given by induction. For $i = n$, (i) and (iii) are identical. Suppose (i) holds for $i = 0,1,\ldots,n$ and (iii) holds for $i = n,n-1,\ldots,k+1$ $(k \geq 0)$. Then it is enough to show that (iii) holds for $i = k$. Expressing $(1-\theta)^{n-i}$ as a power series, (i) may be written for $i = k$ as

$$\sum_{j=0}^{n-k+1} (-1)^j E(\theta^{k+j})\left[\binom{n-k}{j-1} + \binom{n-k}{j}\frac{k+\sqrt{n}/2}{n+\sqrt{n}}\right] = 0$$

By substitution of (iii) with $i = n,n-1,\ldots,k+1$ respectively, the expectations $E(\theta^{n+1}),E(\theta^{n}),\ldots,E(\theta^{k+2})$ are eliminated successively. After m substitutions the highest moment left is $E(\theta^{n-m+1})$ with coefficient

$$\binom{n-k}{n-k-m} \frac{n-m+\sqrt{n}}{n+\sqrt{n}}$$

(as can be proved by induction with respect to m). Hence, after n-k substitutions the summation is reduced to

$$\frac{k+\sqrt{n}}{n+\sqrt{n}} E(\theta^{k+1}) - \frac{k+\sqrt{n}/2}{n+\sqrt{n}} E(\theta^{k}) = 0$$

which is identical to (iii) with $i = k$. □

Theorem 5.15 does not yet provide concrete P-values for which d_m is minimax. To find these it is necessary to consider in more detail the last condition in the theorem: can the vector $(1,\mu_1,\mu_2,\ldots,\mu_{n+1})$, defined by (5.22), be identified as a series of successive moments of some probability distribution on $[1-P,P]$?

A similar question arises in the more general problem of finding minimax rules for the truncated binomial problem. By definition, minimax rules can be derived by maximizing the minimum Bayes risk with respect to the prior distribution. Theorem 5.11 showed that the Bayes risk (5.15) depends on the prior distribution through a limited number of its moments only. Finding minimax rules therefore comes down to maximizing this function, however under the condition that the solution must consist of moments of some prior distribution $\tau \in \Theta^*$. Thus, the central question is, what vectors can be viewed as a series of first successive moments of a probability distribution on a given interval. For an infinite series of moments the answer, due to Hausdorff, is given in the next theorem; it can be found as Theorem 6.6.6 in WIDDER 1971. Compare also WIDDER 1972, Chapter III.

Theorem 5.17. Vector $(c_0,c_1,c_2,\ldots)$ is the vector of moments of some probability distribution on $[0,1]$ if and only if $c_0 = 1$ and

(5.24) $\quad (-\Delta)^k c_n \geq 0$

holds for $k = 0,1,2,\ldots$, with $\Delta c_n := c_{n+1} - c_n$. □

Hence, under the conditions of the theorem,

$$(5.25)\qquad c_i = \int_0^1 u^i dF(u)\ ,\ i = 0,1,2,\ldots$$

holds for some distribution function F on [0,1].

In the next chapter the more complicated problems concerning finite series of moments will be considered. Solutions will be presented in terms of Hankel determinants. Besides, distribution functions will be described that correspond to a given moment vector through (5.25). Since these are not unique, attention will be concentrated on distributions that are as simple as possible in the sense that their support consists of as few points as possible. These results will be used in the final Chapter 7 in order to derive minimax estimators for the symmetric truncated binomial problem.

6. MOMENT PROBLEMS

6.1. Introduction and summary

Consider space $\mathbb{R}_{r+1}$ of all real vectors $c := (c_0, c_1, \ldots, c_r)$ with $r \geq 1$; denote by S_r the linear variety consisting of all vectors $c \in \mathbb{R}_{r+1}$ with first coordinate equal to 1:

$$S_r := \{c \in \mathbb{R}_{r+1} : c_0 = 1\}$$

Let B be any interval in $\mathbb{R}$: closed or (half) open, finite or infinite. Attention will be focused here on the space of all vectors in S_r of which the i-th component equals the i-th moment ($i = 1,2,\ldots,r$) of some probability distribution on B.

Definition 6.1. Let $B \subset \mathbb{R}$ be an interval and let $\mathcal{F}_B$ denote the class of all probability distribution functions F with support in B. Then the space D_B^r consisting of all vectors $c \in S_r$ with the property

$$(6.1) \qquad c_i = \int_B u^i dF(u) \ , \ i = 1,2,\ldots,r$$

for some $F \in \mathcal{F}_B$, is called the moment space (of B). □

So, for any $c \in D_B^r$ at least one distribution exists, restricted to B and with (first) r moments $c_1, c_2, \ldots, c_r$. It is easy to check that D_B^r is convex for any B, and closed if B is compact.

The well-known moment problem, which history traces back to Tchebycheff, is to characterize D_B^r, that is to give (necessary and sufficient) conditions for which $c \in S_r$ belongs to D_B^r. Furthermore, it is of interest to construct an $F \in \mathcal{F}_B$ that through (6.1) corresponds to a given $c \in D_B^r$. Although only the case of finite B will be needed in Chapter 7, infinite intervals B will be discussed as well to attain greater generality.

Section 6.2 presents some results of KARLIN & SHAPLEY 1953 and KARLIN & STUDDEN 1966, characterizing the moment space by means of the nonnegativity of so-called Hankel determinants.

Section 6.3 shows how to find distribution functions that correspond to a given point of the moment space D_B^r. It will be shown that always a step function $F \in \mathcal{F}_B$ can be found. Attention will be concentrated upon step functions with as few steps as possible. The result generalizes the solution for the case $B = \mathbb{R}$ by VON MISES 1964.

The final section 6.4 considers moment vectors corresponding to a symmetric distribution.

6.2. Moment spaces and Hankel determinants

Definition 6.2. For any $c \in S_r$ and any $a,b \in \mathbb{R}$ $(a < b)$ Hankel matrices $\underline{M}_t$, $\bar{M}_t$ and Hankel determinants $\underline{\Delta}_t$, $\bar{\Delta}_t$ $(t = 0,1,\ldots,r)$ are defined as

(6.2) $\underline{M}_{2s} := (c_{i+j})_{i,j=0}^{s}$, $\underline{\Delta}_{2s} := \det(\underline{M}_{2s})$

(6.3) $\bar{M}_{2s} := (-abc_{i+j}+(a+b)c_{i+j+1}-c_{i+j+2})_{i,j=0}^{s-1}$, $\bar{\Delta}_{2s} := \det(\bar{M}_{2s})$

(6.4) $\underline{M}_{2s+1} := (c_{i+j+1}-ac_{i+j})_{i,j=0}^{s}$, $\underline{\Delta}_{2s+1} := \det(\underline{M}_{2s+1})$

(6.5) $\bar{M}_{2s+1} := (bc_{i+j}-c_{i+j+1})_{i,j=0}^{s}$, $\bar{\Delta}_{2s+1} := \det(\bar{M}_{2s+1})$ $\square$

All matrices are symmetric; the suffix of any matrix equals the highest suffix of its elements. An upper (lower) bar indicates the occurrence (absence) of the scalar b. For example,

$$\underline{M}_2 = \begin{pmatrix} c_0 & c_1 \\ c_1 & c_2 \end{pmatrix} \qquad \underline{M}_3 = \begin{pmatrix} c_1-ac_0 & c_2-ac_1 \\ c_2-ac_1 & c_3-ac_2 \end{pmatrix}$$

For any set $S \subset \mathbb{R}_n$, its closure will be denoted by $C\ell\{S\}$, its interior by $\mathrm{Int}\{S\}$ and its boundary by $\mathrm{Bd}\{S\}$.

Theorem 6.3. Let B be any interval in $\mathbb{R}$; take $c \in S_r$. Then $c \in \mathrm{Int}\{D_B^r\}$ if and only if the following Hankel determinants are positive.

B	positive Hankel determinants	
$[a,b]$	$\underline{\Delta}_t$ and $\overline{\Delta}_t$,	$t = 0,1,\ldots,r$
$[a,\infty)$	$\underline{\Delta}_t$,	$t = 0,1,\ldots,r$
$\mathbb{R}$	$\underline{\Delta}_{2t}$,	$t = 0,1,\ldots,\mathrm{ent}(r/2)$

If $c \in \mathrm{Bd}\{D_B^r\}$, then the above determinants are nonnegative. □

A proof of this theorem for $B = [0,1]$ can be found in KARLIN & SHAPLEY 1953 and can easily be generalized to any finite B. The other two cases were proved by KARLIN & STUDDEN 1966. The three main steps of these proofs are outlined below.

(i) Consider vectors y of coefficients, corresponding with polynomials p defined by

$$p(u) = \sum_{i=1}^{r} y_i u^i \ , \ u \in \mathbb{R} \tag{6.6}$$

which have the property of being nonnegative on B:

$$p(u) \geq 0 \ , \ u \in B \tag{6.7}$$

Let $F \in \mathcal{F}_B$ correspond through (6.1) to a given $c \in D_B^r$. Then the obvious statement

$$0 \leq \int_B p(u)dF(u) = \sum_{i=1}^{r} y_i \int_B u^i dF(u) = \sum_{i=1}^{r} y_i c_i$$

implies

$$c \in D_B^r \Rightarrow c^T y \geq 0 \text{ for all } y \tag{6.8}$$

With much more ingenuity the reverse can be shown to hold as well, at least for closed D_B^r.

(ii) Any polynomial p with leading coefficient $y_r \neq 0$ and satisfying (6.6) and (6.7), can be represented by a linear combination of two interlacing polynomials of a special type. For $B = [a,\infty)$ e.g.,

$$p(u) = \alpha \underline{p}_r(u) + \beta \underline{p}_{r-1}(u)$$

where

$$(6.9) \qquad \underline{p}_{2s}(u) = \prod_{j=1}^{s} (u-u_{2j-1})^2$$

$$(6.10) \qquad \underline{p}_{2s+1}(u) = (u-a) \prod_{j=1}^{s} (u-u_{2j})^2$$

with $a \leqslant u_1 \leqslant u_2 \leqslant \ldots \leqslant u_{2s}$. Note that (6.9) and (6.10) are indeed nonnegative on $[a,\infty)$. Consequently, to check the implied statement in (6.8) it suffices to check $c^T y \geqslant 0$ for all vectors of coefficients y, corresponding with polynomials of the type $\underline{p}_t(u)$, $t = 0,1,\ldots,r$.

(iii) A polynomial (6.9) can be written as

$$\sum_{i=0}^{2s} y_i u^i = \underline{p}_{2s}(u) = [\sum_{i=0}^{s} x_i u^i]^2 = \sum_{i=0}^{s} \sum_{j=0}^{s} x_i x_j u^{i+j}$$

which implies

$$(6.11) \qquad c^T y = x^T \underline{M}_{2s} x$$

where $x := (x_0,x_1,\ldots,x_s)$ and $\underline{M}_{2s}$ was defined in (6.2). Now, positiveness of the left-hand side of (6.11) for all y is equivalent to positive definiteness of $\underline{M}_{2s}$, which in turn is equivalent to positiveness of its leading principal minors $\underline{\Delta}_{2t}$, $t = 0,1,\ldots,s$. Similarly,

$$\sum_{i=0}^{2s+1} y_i u^i = \underline{p}_{2s+1}(u) = (u-a)[\sum_{i=0}^{s} x_i u^i]^2 = \sum_{i=0}^{s} \sum_{j=0}^{s} x_i x_j u^{i+j}(u-a)$$

or

$$c^T y = x^T \underline{M}_{2s+1} x$$

leads to positiveness of the Hankel determinants $\underline{\Delta}_{2t+1}$, $t = 0,1,\ldots,s$.

In fact, KARLIN & STUDDEN 1966 studied the more general concept of a Tchebycheff system. A set of continuous functions $f_i : [a,b] \subset \mathbb{R} \to \mathbb{R}$ $(i = 0,1,\ldots,r)$ is called a Tchebycheff system, if

$$\begin{vmatrix} f_0(t_0) & f_0(t_1) & \cdots & f_0(t_r) \\ f_1(t_0) & f_1(t_1) & \cdots & f_1(t_r) \\ \vdots & & & \\ f_r(t_0) & f_r(t_1) & \cdots & f_r(t_r) \end{vmatrix} > 0$$

holds whenever $a \leq t_0 < t_1 < \ldots < t_r \leq b$. The induced 'moment' space is then defined as

$$\left\{c \in S_r : c_i = \int_a^b f_i(u)dF(u),\ i = 0,1,\ldots,r\right\}$$

for some distribution function F on [a,b]. The choice $f_i(u) = u^i$ reduces this situation to the classical moment space considered here.

The proofs outlined above are mainly based on geometric arguments. For the case $r = \infty$, SHOHAT & TAMARKIN 1943 gave a purely algebraic proof. Their more elegant reasoning can be adapted for the case of finite even r. However, for odd r the situation is not clear; for this reason the method will not be introduced here.

6.3. Distribution functions of minimal degree

In this section the problem of how to find a distribution function F on B corresponding with a given point of the moment space D_B^r will be solved. It will be shown that (for finite r) a step function can always be found. Special attention will be paid to step functions which have as few steps as possible. Such a most simple step function is not necessarily unique; uniqueness may be achieved by fixing the location of, for example, the first step. A solution for the case $B = \mathbb{R}$, based on Tchebycheff's findings, was given by VON MISES 1964, Chapter VIII. To simplify the discussion some concepts will be introduced, the most important being the degree of a step function as defined by WALD 1939.

Definition 6.4. Let F be a step function on the interval $B \subset \mathbb{R}$. The degree $d(F)$ of F is defined as the number of different steps, counting a step at an endpoint of B as $\frac{1}{2}$. □

Further, in case some of the Hankel determinants, corresponding with a given interval B through Theorem 6.3, have the value zero, the one with the lowest suffix will be called the minimum zero Hankel determinant. Finally, the class of distribution functions $F \in \mathcal{F}_B$ having moment vector $c \in D_B^r$ will be denoted by $\mathcal{F}_B(c)$.

The case $c \in \mathrm{Bd}\{D_B^r\}$ will be treated first, being the simpler; the case $c \in \mathrm{Int}\{D_B^r\}$ is covered by Theorem 6.7.

Theorem 6.5. For any given $c \in \mathrm{Bd}\{D_B^r\}$, $\mathcal{F}_B(c)$ contains exactly one element. This unique $F \in \mathcal{F}_B(c)$ is a step function with a degree that equals half the index of the minimum zero Hankel determinant.

Proof. For $c \in \mathrm{Bd}\{D_B^r\}$, Theorem 6.3 implies that at least one of the corresponding Hankel determinants has value zero. Assume that the minimum zero Hankel determinant is $\underline{\Delta}_{2s}$, $s \leqslant \mathrm{ent}(r/2)$. Now $\underline{\Delta}_{2s} = 0$ implies the existence of a $(x_0, x_1, \ldots, x_s) \in \mathbb{R}_{s+1}$ with

$$\sum_{i=0}^{s} x_i c_{i+j} = 0 \text{ , } j = 0,1,\ldots,s$$

so that any $F \in \mathcal{F}_B(c)$ must satisfy

$$\int_B [\sum_{i=0}^{s} x_i u^i]^2 dF(u) = \int_B \sum_i \sum_j x_i x_j u^{i+j} dF(u) = \sum_{j=0}^{s} x_j [\sum_{i=0}^{s} x_i c_{i+j}] = 0$$

Since the integrand on the left is nonnegative with at most s different zeros, F is a step function with $d(F) \leq s$. If, however, F has only s-1 steps $u_1, u_2, \ldots, u_{s-1} \in \mathrm{Int}\{B\}$, the relation

$$\int_B u^j \prod_{i=1}^{s-1} (u-u_i)dF(u) = 0, \quad j = 0,1,\ldots,s-1$$

follows, leading to an (excluded) linear dependence between the rows of $\underline{\Delta}_{2(s-1)}$. Similarly, if F has s different steps, but one of them coincides with an endpoint of B, $\underline{\Delta}_{2s-1}$ or $\bar{\Delta}_{2s-1}$ has value 0. So, F has exactly s different steps in $\mathrm{Int}\{B\}$, so that $d(F) = s$. The uniqueness of F easily follows.

In case the minimum zero Hankel determinant is $\bar{\Delta}_{2s}$, a similar reasoning shows that F has s+1 steps at $a = u_1 < u_2 < \ldots < u_s < u_{s+1} = b$, so that again $d(F) = s$. The other two cases are treated in the same way. □

The next theorem considers the actual construction of the single element of $\mathcal{F}_B(c)$. The notations

$$(6.12) \quad \underline{v}_{2s} := (c_{s+j})_{j=0}^{s-1}$$

$$(6.13) \quad \bar{v}_{2s} := (-abc_{s+j-1}+(a+b)c_{s+j}-c_{s+j+1})_{j=0}^{s-2}$$

$$(6.14) \quad \underline{v}_{2s+1} := (c_{s+j+1}-ac_{s+j})_{j=0}^{s-1}$$

$$(6.15) \quad \bar{v}_{2s+1} := (bc_{s+j}-c_{s+j+1})_{j=0}^{s-1}$$

will appear to be useful. Note that $\underline{v}_t$ ($\bar{v}_t$) is the last column of $\underline{M}_t$ ($\bar{M}_t$), leaving out the element in the lower right-hand corner; compare (6.2)-(6.5).

Theorem 6.6. For $c \in \mathrm{Bd}\{D_B^r\}$, let $F \in \mathcal{F}_B(c)$ have k steps in $\mathrm{Int}\{B\}$. Then these steps occur precisely at the zeros of polynomial

$$(6.16) \quad u^k + \sum_{i=0}^{k-1} x_i u^i$$

If the minimum zero Hankel determinant is $\underline{\Delta}_t$, the vector of coefficients $x := (x_0, x_1, \ldots, x_{k-1})$ in (6.16) is the unique solution of

$$(6.17) \quad \underline{M}_{t-2} x = -\underline{v}_t$$

If this determinant is $\bar{\Delta}_t$, x is the unique solution of

$$(6.18) \quad \bar{M}_{t-2} x = -\bar{v}_t$$

Proof. Consider the typical case that the minimum zero determinant is $\underline{\Delta}_{2s}$ for some $s \leqslant \operatorname{ent}(r/2)$. Then the unique $F \in \mathcal{F}_B(c)$ has s steps $u_1, u_2, \ldots, u_s \in \operatorname{Int}\{B\}$. Hence

$$\int_B u^j \prod_{i=1}^{s} (u-u_i) dF(u) = 0, \ j = 0,1,\ldots,s-1$$

By putting

$$\prod_{i=1}^{s} (u-u_i) =: u^s + \sum_{i=0}^{s-1} x_i u^i$$

the system of linear equations

$$\sum_{i=0}^{s-1} x_i c_{i+j} = -c_{s+j}, \ j = 0,1,\ldots,s-1$$

is formed, which can be written as $\underline{M}_{2s-2} x = \underline{v}_{2s}$.

The other cases can be treated similarly. Note that k equals the number of rows of the minimum zero determinant; so $k = \operatorname{ent}(r/2)$ in general, but $k = s-1$ if the minimum zero determinant is $\bar{\Delta}_{2s}$. □

Theorem 6.7. For any $c \in \operatorname{Int}\{B\}$, $\mathcal{F}_B(c)$ contains infinitely many elements. Among them is a step function of (minimal) degree $(r+1)/2$, except in the case $B = \mathbb{R}$ and even r, where a step function of (minimal) degree $(r+2)/2$ exists.

Proof. The first statement immediately follows from the fact that any $c \in \mathrm{Int}\{D_B^r\}$ can be represented in infinitely many ways by a convex combination of boundary points. The same convex combination of the corresponding distribution functions is a distribution function that through (6.1) corresponds with c. To prove the second statement three cases have to be considered, depending on the nature of B.

(i) For $B = [a,b]$ and $c^T = (c_0,c_1,\ldots,c_r) \in \mathrm{Int}\{D_B^r\}$, define $c_* \in \mathrm{Bd}\{D_B^{r+1}\}$ by $c_*^T := (c_0,c_1,\ldots,c_r,c_{r+1})$, where c_{r+1} is the solution of the equation $\underline{\Delta}_{r+1} = 0$. Then Theorem 6.5 ensures the existence of a unique step function $F_* \in \mathcal{F}_B(c_*)$ with $d(F_*) = (r+1)/2$. Of course, F_* belongs to $\mathcal{F}_B(c)$ as well. Similarly, a second distribution function F^* of (minimal) degree $(r+1)/2$ is obtained by means of $c^* \in \mathrm{Bd}\{D_B^{r+1}\}$ having as its last coordinate the solution of $\bar{\Delta}_{r+1} = 0$. Note that exactly one of these two distribution functions has its first step in a, as follows from the construction in Theorem 6.6.

(ii) For $B = [a,\infty)$, a distribution function of minimal degree can be found only by means of c_* introduced above, and is unique by consequence. Note that for even r the first step occurs in a automatically.

(iii) For $B = \mathbb{R}$, a similar argument shows that for odd r a unique $F \in \mathcal{F}_B(c)$ with degree $(r+1)/2$ exists. For $r = 2m$, the points $(c_0,c_1,\ldots,c_{2m},c_{2m+1},c_{2m+2}) \in D_{\mathbb{R}}^{r+2}$ will be considered. For any c_{2m+1}, $\underline{\Delta}_{2m+2} = 0$ determines a boundary point of $D_{\mathbb{R}}^{r+2}$; hence, infinitely many $F \in \mathcal{F}_{\mathbb{R}}(c)$ with $d(F) = (r+2)/2$ exist. Note that such an F is uniquely determined by its first step; if this first step occurs in, say, $a \in \mathbb{R}$, $c \in D_{[a,\infty)}^r$ as well. □

The construction of a distribution function of minimal degree corresponding with a given $c \in \mathrm{Int}\{D_B^r\}$ easily follows and starts by equating to zero one of the determinants $\underline{\Delta}_{r+1}$, $\bar{\Delta}_{r+1}$ or $\underline{\Delta}_{r+2}$, depending on the nature of B. The last case, referring to $B = \mathbb{R}$, and even r, deserves the choice of an arbitrary c_{r+1}. The details follow from Theorem 6.6.

All elements of $\mathcal{F}_B(c)$ must show certain intersection properties, particularly with respect to an F of minimal degree. This follows from the next two lemmas, due to VON MISES 1964, Chapter VIII.

Lemma 6.8. If F_1 and F_2 both belong to $\mathcal{F}_B(c)$ for some $c \in \mathrm{Int}\{D_B^r\}$, then F_2-F_1 changes sign on B at least r times.

Proof. The relations

$$\int_B u^i dF_2(u) = c_i = \int_B u^i dF_1(u), \quad i = 0,1,\ldots,r$$

imply that for any r-th degree polynomial $p_r(u)$

$$\int_B p_r(u) d[F_2(u) - F_1(u)] = 0$$

or, by partial integration,

$$\int_B [F_2(u) - F_1(u)] p_r'(u) du = 0$$

Now, if F_2-F_1 should change sign only in the r-1 points $u_1,u_2,\ldots,u_{r-1}$, choosing $p_r'(u) = \Pi_i(u-u_i)$ would make the integrand never change sign on B, which provides a contradiction. □

Lemma 6.9. If F_1 and F_2 both belong to $\mathcal{F}_B$ with $d(F_1) = m$, then F_2-F_1 changes sign at most 2m-1 times.

Proof. The lemma follows at once from the observation that nondecreasing function F_2 can intersect each horizontal and vertical segment of the graph of F_1 only once. □

Now consider $c \in \mathrm{Int}\{D_B^r\}$ and let $F_0 \in \mathcal{F}_B(c)$ have minimal degree. If $d(F_0) = (r+1)/2$, the foregoing lemmas imply, that any $F \in \mathcal{F}_B(c)$ (not identical to F_0) intersects F_0 the maximum number of times, i.e. r. Denoting by S(F) the support of the distribution determined by F, it follows that

(6.19) $\quad \inf S(F_0) \geq \inf S(F), \quad \sup S(F_0) \leq \sup S(F)$

holds for all $F \in \mathcal{F}_B(c)$. Equality can occur only if F_0 has a step at an endpoint of B. Only in the case $B = \mathbb{R}$ and r even, $d(F_0)$ equals $(r+2)/2$; the number of intersections between F_0 and any $F \in \mathcal{F}_B(c)$ is then either r or r+1.

Figure 6.10 presents a sketch of the two distribution functions of minimal degree corresponding with some $c \in \mathrm{Int}\{D_B^r\}$, where $B = [a,b]$; separate pictures are given for even and odd r respectively. Indeed, the degree is $(r+1)/2$, in agreement with Theorem 6.7; the number of intersections is r, which agrees with the last paragraph.

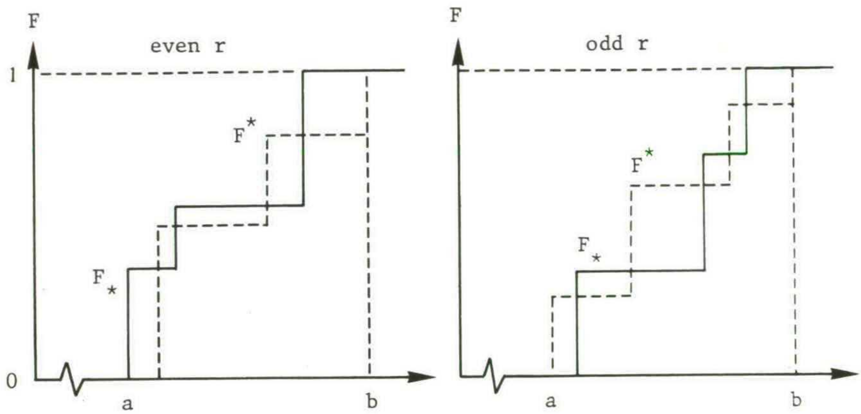


Figure 6.10. Distribution functions of minimal degree for r = 4 and 5

The case $B = [a,\infty)$ can be illustrated too by means of this figure; the distribution functions F^* disappear from the scene (and of course, b has no longer any meaning).

6.4. Moment vectors of symmetric distributions

Now the question will be answered which moment vectors in D_B^r correspond with a symmetric distribution on B. Only finite B will be considered here since this is the only case needed. KARLIN & SHAPLEY 1953 touched on this question for the case $B = [0,1]$. Their geometric approach in Theorem 25.6 shows that the condition

(6.20) $\underline{\Delta}_{2s+1} = \bar{\Delta}_{2s+1}$, $s = 0,1,\ldots,\text{ent}((r-1)/2)$

is necessary. Here, it will be shown by a direct argument, that (6.20) is necessary as well as sufficient for any finite B. The proof is given for a specific B and can be extended to general finite intervals.

Theorem 6.11. Let $B = [-1,1]$ and take any $c \in D_B^r$. Then $\mathcal{F}_B(c)$ contains a symmetric distribution with respect to zero if and only if (6.20) holds.
Proof. The special choice of B reduces the Hankel matrices in (6.4) and (6.5) to

(6.21) $\underline{M}_{2s+1} = (c_{i+j}+c_{i+j+1})_{i,j=0}^{s}$

(6.22) $\bar{M}_{2s+1} = (c_{i+j}-c_{i+j+1})_{i,j=0}^{s}$

To show the 'only if' part, assume that $\mathcal{F}_B(c)$ contains a distribution that is symmetric with respect to zero. It follows that all odd moments vanish, further simplifying (6.21) and (6.22) to

$$\underline{M}_{2s+1} = (c_{2\text{ent}((i+j+1)/2)})_{i,j=0}^{s}$$

$$\bar{M}_{2s+1} = ((-1)^{i+j}c_{2\text{ent}((i+j+1)/2)})_{i,j=0}^{s}$$

By changing in $\bar{\Delta}_{2s+1}$ the signs of the even columns and, after that, the signs of the even rows, $\underline{\Delta}_{2s+1}$ is obtained; so, (6.20) holds indeed.

The 'if' part of the statement will be proved by showing that equality of the determinants of (6.21) and (6.22) for $s = 0,1,\ldots,k$ implies $c_{2s+1} = 0$, $s = 0,1,\ldots,k$. Use induction with respect to k; for $k = 0$ the statement trivially holds. Next, let the statement hold for $s = 0,1,\ldots,k-1$ and assume $\underline{\Delta}_{2s+1} = \bar{\Delta}_{2s+1}$ for $s = 0,1,\ldots,k$. Since $c_{2s+1} = 0$ for $s = 0,1,\ldots,k-1$, changing signs in the even columns and, after that, in the even rows, now leads to the equality

$$0 = \underline{\Delta}_{2k+1} - \bar{\Delta}_{2k+1} = 2c_{2k+1}\underline{\Delta}_{2k-1}$$

If $\underline{\Delta}_{2k-1} \neq 0$, the induction argument is completed; hence, assume $\underline{\Delta}_{2k-1} = 0$. Then $\tilde{c} := (c_0, c_1, \ldots, c_{2k}) \in \mathrm{Bd}\{D_B^{2k}\}$ according to Theorem 6.3, and Theorem 6.5 implies the existence of a unique $F \in \mathcal{F}_B(\tilde{c})$. However, the function F^* defined by

$$F^*(x) = [1 + F(x) - F(-x^-)]/2$$

is a (symmetric) distribution function in $\mathcal{F}_B(\tilde{c})$ as well. It follows that $F = F^*$ so that $c_{2k+1} = 0$, which completes the proof. □

In the case of a general finite closed interval $B = [a,b]$, let X be a random variable with support in B and moment vector c. Then $Y := (2X-a-b)/(b-a)$ defines a random variable with support in $[-1,1]$; denote the corresponding moment vector by $\tilde{c}$. Now it can be shown that (6.20) holds for c and $[a,b]$ if and only if (6.20) holds for $\tilde{c}$ and $[-1,1]$. Since X and Y are simultaneously symmetric, Theorem 6.11 is valid for general intervals $[a,b]$. The detailed proof can be found in MOORS 1979, Theorem 6.3.

Next, consider the construction of a symmetric $F \in \mathcal{F}_B(c)$ of minimal degree, assuming that (6.20) is satisfied. For $c \in \mathrm{Bd}\{D_B^r\}$, F is unique; hence, by the construction outlined in the proof of Theorem 6.6 a symmetric F is obtained. For $c \in \mathrm{Int}\{D_B^r\}$ and $B = [a,b]$, case (i) of the proof of Theorem 6.7 applies. For odd r, the construction presented there again leads to symmetric F_* and F^*. However, for even r, equating $\underline{\Delta}_{r+1}$ or $\overline{\Delta}_{r+1}$ to zero contradicts (6.20) and leads to two asymmetric distributions. Now, two symmetric distributions of minimal degree can be obtained by imposing $\underline{\Delta}_{r+1} = \overline{\Delta}_{r+1}$ and equating to zero either $\underline{\Delta}_{r+2}$ or $\overline{\Delta}_{r+2}$.

The final theorem of this chapter summarizes these results.

Theorem 6.12. Let B be a closed finite interval in $\mathbb{R}$, and take any $c \in D_B^r$. Then $\mathcal{F}_B(c)$ contains a symmetric distribution F with respect to the midpoint of B if and only if (6.20) holds. For $c \in \mathrm{Int}\{D_B^r\}$ the minimal degree of F is at most $(r+1)/2$ for odd r and $(r+2)/2$ for even r. □

Some detailed examples of the theory in this chapter will be presented in the next chapter, particularly in Section 7.2.

7. MINIMAX ESTIMATORS

7.1. Introduction and summary

After the preparatory work in Sections 5.4 and 5.5 and in Chapter 6, minimax estimators can be derived for the symmetric truncated binomial problem of Definition 5.1 with h the identity. The method to be followed is based on Theorem 1.11 and is therefore equivalent to the construction of a least favorable prior distribution. The starting point is the minimum Bayes risk $r(\tau,d_\tau)$ of a prior distribution $\tau \in \Theta^*$, defined by (1.8); the notation will be abbreviated:

$$(1.8) \qquad r_\tau := r(\tau,d_\tau) = \inf_{d \in D} r(\tau,d)$$

The concavity of r_τ will appear to be of interest.

Lemma 7.1. The minimum Bayes risk (1.8) is concave in τ.
Proof. Take two prior distributions τ and $\tau' \in \Theta^*$ and define $\tau'' := \lambda\tau + (1-\lambda)\tau'$ for any $\lambda \in (0,1)$; then $\tau'' \in \Theta^*$ as well. It follows at once that $r(\tau,d)$, the expected risk if θ has distribution τ, satisfies

$$r(\tau'',d) = \lambda r(\tau,d) + (1-\lambda)r(\tau',d)$$

Taking the infimum with respect to all $d \in D$ proves the lemma. □

Now, the problem is to maximize r_τ with respect to all $\tau \in \Theta^*$. Two approaches present themselves.

(i) From (5.11), (5.12) and (5.15) it follows that r_τ depends on τ only through moment vector $\mu := (\mu_0,\mu_1,\ldots,\mu_{n+1})$, where n denotes the number of observations. While maximizing $r_\tau(\mu)$ all vectors μ must be considered that belong to the moment space D^r_B, where now $r = n+1$ and B is the finite closed interval $[1-P,P]$. Characterizations of moment spaces by means of Hankel determinants were given in Theorem 6.3 and it was noted before that D^r_B is convex and closed. Combination of these elements implies that derivation of a minimax estimator can be reduced to a convex programming problem.

Theorem 7.2. Deriving a minimax estimator for the symmetric truncated binomial problem with h the identity is equivalent to the following convex programming problem: minimize the convex function $-r_\tau(c)$ where $c := (c_0, c_1, \ldots, c_{n+1})$, subject to the conditions

$$(7.1) \qquad \underline{\Delta}_t \geq 0 \ , \ \bar{\Delta}_t \geq 0 \ , \ t = 0,1,\ldots,n+1$$

which describe a convex space in $\mathbb{R}_{n+1}$. □

(ii) For any point of moment space D_B^r, Section 6.3 showed the existence of a corresponding distribution function F with $d(F) \leq (r+1)/2$. The number of steps of F is therefore at most $\text{ent}((r+3)/2)$. Denote for the truncated binomial problem the point masses by y_i and their locations by x_i $(i = 1,2,\ldots,p)$, where now

$$(7.2) \qquad p := \text{ent}((n+4)/2)$$

Then r_τ can be viewed as a function of

$$w := (x_1, x_2, \ldots, x_p, y_1, y_2, \ldots, y_p)$$

and has to be maximized with respect to w.

Theorem 7.3. Deriving a minimax rule for the symmetric truncated binomial problem with h the identity is equivalent to the following non-linear programming problem: maximize $r_\tau(w)$ subject to the linear constraints

$$(7.3) \qquad \begin{cases} 1-P \leq x_i \leq P \ , \ i = 1,2,\ldots,p \\ 0 \leq y_i \leq 1 \ , \quad i = 1,2,\ldots,p \\ \sum_{i=1}^{p} y_i = 1 \end{cases} \quad \square$$

Note that in many cases the above number of p steps is larger than necessary; in case of a unique least favorable prior, some of the locations x_i wil coincide or have point mass zero. Especially for $c \in \mathrm{Bd}\{D_B^r\}$ therefore, a smaller number of point masses than indicated by (7.2) can be chosen in advance.

A further simplification is possible, using the invariance of the present estimation problem. Theorems 1.12 and 1.13 imply that all attention may be confined to minimax estimators and least favorable distributions that are symmetric with respect to $\frac{1}{2}$. Theorem 6.12 gave the additional condition (6.20) for moment vectors corresponding with a symmetric distribution and can be used to adapt Theorem 7.2; as a result Theorem 7.2 remains valid when the conditions

$$(7.4) \qquad \underline{\Delta}_{2s+1} = \overline{\Delta}_{2s+1} \;, \; s = 0,1,\ldots,\mathrm{ent}(n/2)$$

are added to the programming problem. At the same time, Theorem 6.12 showed the existence of symmetric distributions with a degree of at most $\mathrm{ent}((r+2)/2)$. Such a distribution is fully determined by the k locations x_i smaller than $\frac{1}{2}$, the corresponding point masses y_i and the (possible) point mass y_{k+1} at $\frac{1}{2}$. The construction of F_* in Theorem 6.6 shows that $k = \mathrm{ent}((r+2)/4)$ suffices.

Consequently, Theorem 7.3 remains valid when condition (7.3) is replaced by

$$(7.5) \qquad \begin{cases} 1-P \leqslant x_i < \frac{1}{2} \;, & i = 1,2,\ldots,m \\ 0 \leqslant y_i \leqslant 1 \;, & i = 1,2,\ldots,m+1 \\ 2 \sum_{i=1}^{m} y_i + y_{m+1} = 1 & \end{cases}$$

where y_{m+1} denotes the point mass at $\frac{1}{2}$ and

$$(7.6) \qquad m := \mathrm{ent}((n+3)/4)$$

Again, several locations may coincide or have point mass zero.

The general method for calculating minimax estimators, presented by NELSON 1966, offers a third approach towards the solution of the present problem. The method proceeds iteratively according to the following principle. Let τ_k be the prior distribution obtained after the k-th iteration; the corresponding Bayes rule d_k has risk function R_k. If R_k attains its maximum in $\theta_{k+1} \in \Theta$, a new prior is constructed with an additional point mass in θ_{k+1}. Hence, each iteration may extend the number of steps, so that no allowance is made for the maximum number of steps needed for the symmetric truncated binomial problem. It is feared that this will unfavorably influence the computer time needed; even more important is the fact that least favorable priors of minimal degree are of interest by themselves.

Method (i) was applied first, using an Algol 68 program, which was based upon Powell's quasi-Newtonian algorithm (to be discussed in more detail in Section 7.3). To be short: although the program produced the correct results for very small n, it failed for n as low as 4 and P anywhere near $\frac{1}{2}$. It is believed that this failure is caused by the highly non-linear character of the constraints.

Next, method (ii) was implemented and worked nicely. A detailed description is given in Section 7.3, where the numerical results are presented as well. First, however, the cases n=1 up to 3 are analytically solved in Section 7.2, at the same time offering detailed examples of the theory in Chapter 6.

The final Section 7.4 numerically presents minimax estimators for the symmetric truncated binomial problem, now, however, with the weighted quadratic loss function $(\theta-a)^2/\{\theta(1-\theta)\}$.

7.2. Small sample sizes

Minimax estimators for the parameter θ of the symmetric truncated binomial problem will now be derived for n=1 up to 3, successively. The main results are summarized in Table 7.9. A similar analysis is possible for n=4, but is omitted here. For easy reference, the Hankel determinants with the highest suffix that will be needed in this section are written out below for $B = [1-P,P]$. Of course, the ones with lower suffix can be found in the upper left-hand corners.

$$\underline{\Delta}_4 = \begin{vmatrix} c_0 & c_1 & c_2 \\ c_1 & c_2 & c_3 \\ c_2 & c_3 & c_4 \end{vmatrix} \qquad \bar{\Delta}_4 = \begin{vmatrix} -P(1-P)c_0 + c_1 - c_2 & -P(1-P)c_1 + c_2 - c_3 \\ -P(1-P)c_1 + c_2 - c_3 & -P(1-P)c_2 + c_3 - c_4 \end{vmatrix}$$

$$\bar{\Delta}_5 = \begin{vmatrix} Pc_0 - c_1 & Pc_1 - c_2 & Pc_2 - c_3 \\ Pc_1 - c_2 & Pc_2 - c_3 & Pc_3 - c_4 \\ Pc_2 - c_3 & Pc_3 - c_4 & Pc_4 - c_5 \end{vmatrix}$$

$$\underline{\Delta}_5 = \begin{vmatrix} c_1 - (1-P)c_0 & c_2 - (1-P)c_1 & c_3 - (1-P)c_2 \\ c_2 - (1-P)c_1 & c_3 - (1-P)c_2 & c_4 - (1-P)c_3 \\ c_3 - (1-P)c_2 & c_4 - (1-P)c_3 & c_5 - (1-P)c_4 \end{vmatrix}$$

For a distribution function F with k steps y_i in the locations x_i the compact notation

$$F = \begin{Bmatrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \end{Bmatrix}$$

will be used, where the x_i are in ascending order. If F is symmetric, this notation can be abbreviated still further:

$$F = \begin{Bmatrix} x_1 & x_2 & \cdots & x_{\mathrm{ent}(k/2)} \\ y_1 & y_2 & \cdots & y_{\mathrm{ent}(k/2)} \end{Bmatrix}_S$$

where the symbol S is used as a reminder. Note that now all x_i are to the left of the point of symmetry. If the y_i do not sum to $\frac{1}{2}$, the remainder is located at the point of symmetry.

Example 7.4. For n=1 symmetry conditions (7.4) reduce to $\underline{\Delta}_1 = \overline{\Delta}_1$, or $c_1 = \frac{1}{2}$, while (5.15) becomes

$$r(\tau,d_\tau) = \mu_2 - \frac{(\mu_1-\mu_2)^2}{1-\mu_1} - \frac{\mu_2^2}{\mu_1}$$

The problem of Theorem 7.2 can now be simplified to: maximize

$$(7.7) \qquad r_\tau(c) = (-8c_2^2+6c_2-1)/2$$

subject to the condition that $\underline{\Delta}_2$ and $\overline{\Delta}_2$ are nonnegative. This condition is equivalent to

$$(7.8) \qquad \tfrac{1}{4} \leqslant c_2 \leqslant \tfrac{1}{2} - P(1-P) = (1-\phi)/4$$

where, as previously,

$$(7.9) \qquad \phi := (2P-1)^2$$

The unconstrained maximum is attained for $c_2 = 3/8$, but moment vector $(1,\frac{1}{2},3/8)$ satisfies the constraints only for $\phi \geqslant \frac{1}{2}$. For $0 < \phi \leqslant \frac{1}{2}$, the problem is solved by moment vector $(1,\frac{1}{2},(1+\phi)/4)$, which is on the boundary of the moment space. The maximum of r_τ follows, while the minimax estimator is found from (5.14). The results are in full agreement with Example 2.12.

In addition, a least favorable distribution function F of minimal degree can be found now. Case $\phi \leqslant \frac{1}{2}$ is very simple: the minimum zero determinant is $\overline{\Delta}_2$, hence Theorem 6.6 implies that the unique F has no steps in the interior of [1-P,P]. It follows that F has two steps, in 1-P and P respectively, of size $\frac{1}{2}$ each:

$$(7.10) \qquad F = \left\{ \begin{matrix} 1-P \\ \\ \frac{1}{2} \end{matrix} \right\}_S$$

In case $\phi > \frac{1}{2}$, moment vector $(1,\frac{1}{2},3/8)$ maximizing r_τ is in the interior of moment space $D^2_{[1-P,P]}$. Several distribution functions corresponding with this vector will be constructed now with the aid of Theorems 6.7 and 6.12.

(i) Take $\underline{\Delta}_3 = 0$ as additional constraint to derive F_* with degree 3/2. Then (6.17) reduces to $\underline{M}_1 x = -\underline{v}_3$ or, equivalently, to $-x = (P-\frac{1}{4})/(2P-1)$. The only step of F_* inside $(1-P,P)$ is in $-x$, while the second step is located in $1-P$; by consequence,

$$F_* = \begin{Bmatrix} 1-P & (P-\frac{1}{4})/(2P-1) \\ \\ 2\phi/(2\phi+1) & 1/(2\phi+1) \end{Bmatrix}$$

(ii) The choice $\bar{\Delta}_3 = 0$ can be analysed in the same way. However, symmetry considerations at once give

$$F^* = \begin{Bmatrix} (P-\frac{3}{4})/(2P-1) & P \\ \\ 1/(2\phi+1) & 2\phi/(2\phi+1) \end{Bmatrix}$$

(iii) F_* and F^*, obtained above, are not symmetric. A symmetric distribution function is $(F_*+F^*)/2$, but a lower degree can be attained by taking $\underline{\Delta}_3 = \bar{\Delta}_3$ and either $\underline{\Delta}_4 = 0$ or $\bar{\Delta}_4 = 0$ (see Theorem 6.12). Symmetry relation $\underline{\Delta}_3 = \bar{\Delta}_3$ gives

$$(7.11) \qquad c_3 = (6c_2-1)/4$$

The solution of $\underline{M}_2 x = -\underline{v}_4$ is $x = (1/8,-1)$ and $\bar{M}_2 x = -\bar{v}_4$ gives $x = \frac{1}{2}$. The two resulting symmetric distribution functions of degree 2, to be denoted by $\underline{F}$ and $\bar{F}$, are

$$\underline{F} = \begin{Bmatrix} 1/2 - 1/\sqrt{8} \\ \\ 1/2 \end{Bmatrix}_S \qquad \bar{F} = \begin{Bmatrix} 1-P \\ \\ 1/(4\phi) \end{Bmatrix}_S$$

The four least favorable prior distributions derived here for $\phi > \frac{1}{2}$ are drawn in Figure 7.5. □

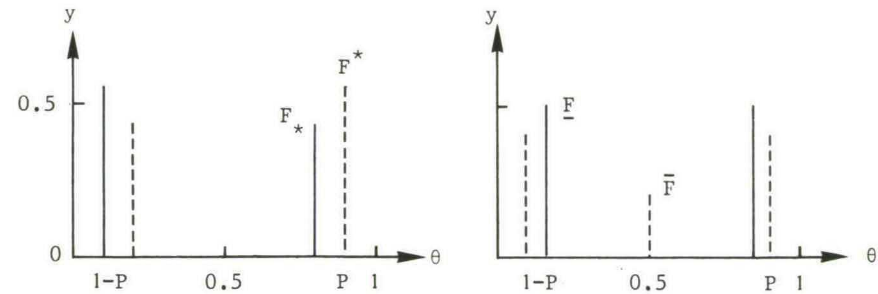


Figure 7.5. Least favorable distributions for the symmetric truncated binomial problem with n=1 (case P = 0.9)

Example 7.6. Case n=2 very much resembles the preceding case. Symmetry conditions (7.4) are equivalent to $c_1 = \frac{1}{2}$ and (7.11); nonnegativity conditions (7.1) again can be reduced to (7.8). The minimum Bayes risk now reads

$$(7.12) \qquad r_\tau(c) = (-8c_2^2+6c_2-1)/(8c_2)$$

The solution of the problem in Theorem 7.2 supplemented by (7.4), reads

$$c_2 = \begin{cases} (1+\phi)/4 & \text{for } \phi \leq \sqrt{2}-1 \\ 1/\sqrt{8} & \text{for } \phi \geq \sqrt{2}-1 \end{cases}$$

In the former case, the unique least favorable distribution appears to be (7.10) again. For $\phi \geq \sqrt{2}-1$, putting $\underline{\Delta}_4$ equal to zero leads to

$$F_* = \begin{Bmatrix} 1/2 - \sqrt{\sqrt{2}-1}/2 \\ 1/2 \end{Bmatrix}_S$$

And $\bar{\Delta}_4 = 0$ gives

$$F^* = \begin{Bmatrix} 1-P \\ \\ (\sqrt{2}-1)/(2\phi) \end{Bmatrix}_S$$

Both are symmetric and have (minimal) degree 2. □

Example 7.7. For n=3 nonnegativity condition (7.8) is augmented by $\underline{\Delta}_4 \geq 0$ and $\bar{\Delta}_4 \geq 0$ or, equivalently,

$$(7.13) \quad c_2^2 + c_2 - \tfrac{1}{4} \leq c_4 \leq \frac{\phi+6}{4} c_2 - \frac{\phi+5}{16}$$

Derivation of $r_\tau(c)$ from (5.15) is rather cumbersome; it is rewarding to use the 'central moments' m_i, defined by

$$(7.14) \quad m_i := E_\tau(2\theta-1)^i$$

so that by symmetry $m_1 = m_3 = 0$ and

$$m_2 = 4c_2 - 1 \ , \ m_4 = 16c_4 - 24c_2 + 5$$

The problem of Theorem 7.2 can now be reformulated as: maximize

$$(7.15) \quad r_\tau(m_2,m_4) = \frac{6m_2^2m_4 - 2m_2m_4^2 - 3m_2^3 - m_4^2 - m_2^2 + m_2}{4(1-m_2)(1+3m_2)}$$

subject to the condition

$$(7.16) \quad m_2^2 \leq m_4 \leq \phi m_2$$

Figure 7.8 shows some curves where (7.15) is constant, as well as the feasible set defined by (7.16).

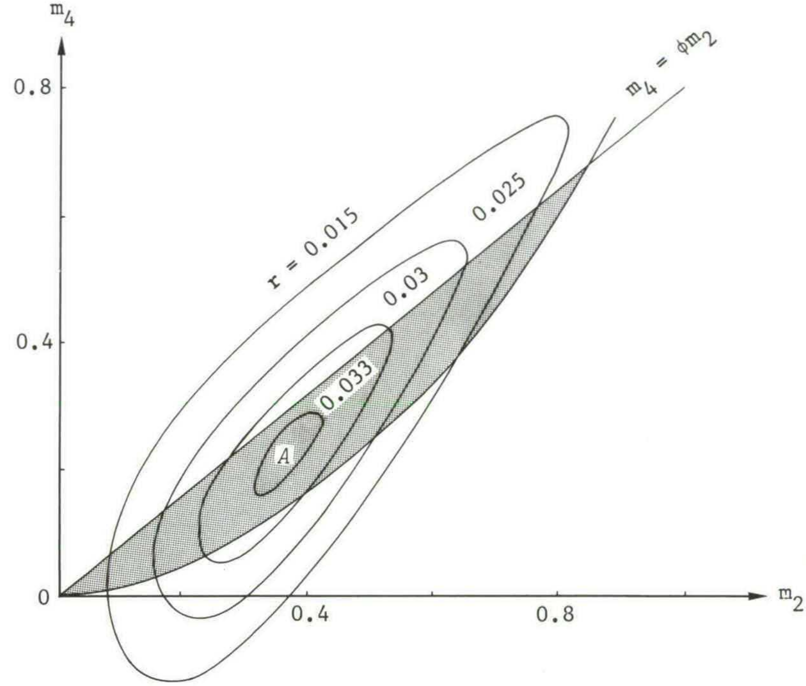


Figure 7.8. Isohypses of r_τ for n=3 (ϕ = 0.8)

To solve this problem three different intervals of ϕ-values have to be distinguished.

(i) The unconstrained maximum of (7.15) equals $(2-\sqrt{3})/8$; it is attained for

$$m_2 = (\sqrt{3}-1)/2 \ , \ m_4 = (2\sqrt{3}-3)/2$$

and lies in the feasible set for $\phi \geqslant (3-\sqrt{3})/2 = 0.634$.

For smaller values of ϕ the solution will be on line L defined by $m_4 = \phi m_2$; substitution gives the values $r_\tau^{(L)}$ of (7.15) on L:

$$(7.17) \qquad r_\tau^{(L)}(m_2) = \frac{m_2^3 f(\phi) - m_2^2(\phi^2+1) + m_2}{4(1-m_2)(1+3m_2)}$$

where

$$(7.18) \quad f(\phi) := -2\phi^2 + 6\phi - 3$$

The derivative of (7.17) with respect to m_2 equals

$$(7.19) \quad \frac{-3m_2^4 f(\phi)+4m_2^3 f(\phi)+2m_2^2[2f(\phi)-3\phi+2]-2m_2(\phi^2+1)+1}{4(1-m_2)^2(1+3m_2)^2}$$

(ii) The solution coincides with the 'top' of the feasible set ($m_2 = \phi$, $m_4 = \phi^2$), if and only if (7.19) is nonnegative for $m_2 = \phi$; this is equivalent to the inequality

$$6\phi^4 - 14\phi^3 - 9\phi^2 + 1 \geqslant 0$$

or, equivalently, $0 \leqslant \phi \leqslant 0.283$.

(iii) For $0.283 \leqslant \phi \leqslant 0.634$ the problem is solved numerically by equating (7.19) to zero. Some values obtained are

ϕ	0.33	0.43	0.53	0.63
m_2	0.29328	0.31634	0.34074	0.36510
$r_\tau^{(L)}(m_2)$	0.03136	0.03241	0.03318	0.03349

The unique symmetric least favorable prior distribution in case (ii) is given by (7.10), while no closed expression exists in case (iii). In case (i), with some effort the two symmetric least favorable priors of degree 3 can be found:

$$\underline{F} = \begin{Bmatrix} 1/2 - \sqrt{(3-\sqrt{3})/8} \\ \\ 1/(2\sqrt{3}) \end{Bmatrix}_S \qquad \bar{F} = \begin{Bmatrix} 1-P & \frac{1}{2}-a \\ \\ y_1 & \frac{1}{2}-y_1 \end{Bmatrix}_S$$

where

$$(7.20) \quad y_1 = \frac{(3\sqrt{3}-5)/2}{2\phi^2-2(\sqrt{3}-1)\phi+2\sqrt{3}-3}, \quad a^2 = \frac{\sqrt{3}-1}{8}\,\frac{2\phi-3+\sqrt{3}}{2\phi+1-\sqrt{3}}$$

Note that $\underline{F}$ has one step located below $\frac{1}{2}$, in accordance with (7.6). $\square$

Table 7.9. Minimax estimators d_m and least favorable priors for the symmetric truncated binomial problem; small sample sizes

n	ϕ	c_2	c_3	c_4	$d_m(n)$	$d_m(n-1)$	minimax value	symmetric least favorable priors with degree $\le$ ent[(n+3)/2]
1	≤ 0.5	$\frac{\phi+1}{4}$			$\frac{\phi+1}{2}$		$\frac{\phi(1-\phi)}{4}$	$\begin{Bmatrix} 1-P \\ 1/2 \end{Bmatrix}_S$
	≥ 0.5	$\frac{3}{8}$			$\frac{3}{4}$		$\frac{1}{16}$	$\begin{Bmatrix} (1-1/\sqrt{2})/2 \\ 1/2 \end{Bmatrix}_S$ $\begin{Bmatrix} 1-P \\ 1/(4\phi) \end{Bmatrix}_S$
2	≤ 0.414	$\frac{\phi+1}{4}$	$\frac{3\phi+1}{8}$		$\frac{3\phi+1}{2(\phi+1)}$		$\frac{\phi(1-\phi)}{4(\phi+1)}$	$\begin{Bmatrix} 1-P \\ 1/2 \end{Bmatrix}_S$
	≥ 0.414	$\frac{1}{\sqrt{8}}$	$\frac{3\sqrt{2}-2}{8}$		$\frac{3-\sqrt{2}}{2}$		$\frac{3-2\sqrt{2}}{4}$	$\begin{Bmatrix} (1-\sqrt{\sqrt{2}-1})/2 \\ 1/2 \end{Bmatrix}_S$ $\begin{Bmatrix} 1-P \\ (\sqrt{2}-1)/(2\phi) \end{Bmatrix}_S$
3	≤ 0.283	$\frac{\phi+1}{4}$	$\frac{3\phi+1}{8}$	$\frac{\phi^2+6\phi+1}{16}$	$\frac{\phi^2+6\phi+1}{2(3\phi+1)}$	$\frac{\phi+1}{2}$	$\frac{\phi(1-\phi)^2(2\phi+1)}{4(3\phi+1)}$	$\begin{Bmatrix} 1-P \\ 1/2 \end{Bmatrix}_S$
	≥ 0.634	$\frac{\sqrt{3}+1}{8}$	$\frac{3\sqrt{3}-1}{16}$	$\frac{8\sqrt{3}-7}{32}$	$(5-\sqrt{3})/4$	$\frac{9-\sqrt{3}}{12}$	$\frac{2-\sqrt{3}}{8}$	$\begin{Bmatrix} (1-\sqrt{(3-\sqrt{3})/2})/2 \\ 1/(2\sqrt{3}) \end{Bmatrix}_S$

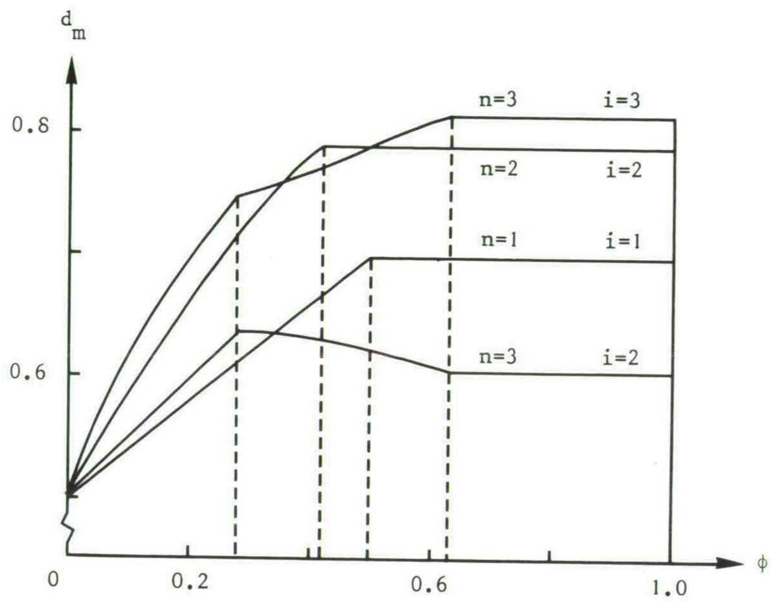


Figure 7.10. Minimax rules for the symmetric truncated binomial problem; small sample sizes

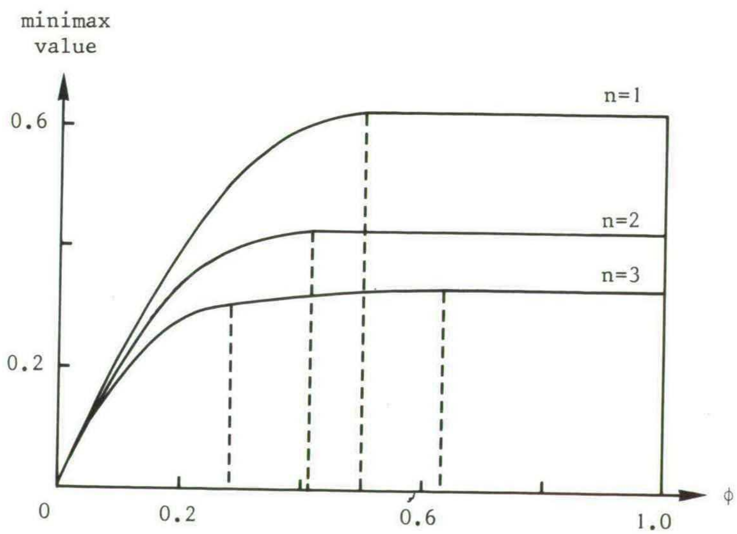


Figure 7.11. Minimax values for the symmetric truncated binomial problem; small sample sizes

Table 7.9 summarizes the main findings of the three preceding examples. Figures 7.10 and 7.11, respectively, show the behavior of the minimax estimator and minimax value $\max_\tau r_\tau$. Note the curious behavior of the minimax rule for n=3 and i=2: for moderate values of ϕ it is decreasing in ϕ. This means that if the parameter space [1-P,P] widens, observing two successes in three experiments leads to lower estimates of θ.

7.3. General sample sizes

First of all, the computer program will be described that was used to calculate the minimax rules starting from Theorem 7.3. The body of the program is a general algorithm for solving constrained nonlinear minimization problems, where the constraints may be linear or nonlinear, equalities or inequalities.

Definition 7.12. Let $x \in \mathbb{R}_n$, $F : \mathbb{R}_n \to \mathbb{R}$ and $c := (c_1, c_2, \ldots, c_m)$ with $c_i : \mathbb{R}_n \to \mathbb{R}$ for $i = 1,2,\ldots,m$; it is assumed that all these functions have continuous second partial derivatives. A problem that can be stated as: minimize F(x) subject to the constraints

$$(7.21) \qquad c_i(x) \begin{cases} = 0 \text{ for } i = 1,2,\ldots,m' \\ \geq 0 \text{ for } i = m'+1, m'+2, \ldots, m \end{cases}$$

is called a nonlinear programming problem. □

A fast algorithm to solve this problem was given by POWELL 1978. It concerns a quasi-Newtonian method, meaning that no analytical expression for the Hessian of F is needed; instead, H(F) is approximated numerically. The method was implemented by VAN DEN AKER 1980 for use on the ICL 2966 computer of Tilburg University. The algorithm is outlined below by presenting the four main steps of any iteration. Until stated otherwise entities μ, x and B have values which were obtained in the foregoing iteration by means of formulas (7.25), (7.28) and (7.29) respectively.

(i) Solve the following linear programming problem in $d \in \mathbb{R}_n$ and $\varepsilon \in \mathbb{R}$: maximize ε subject to the linear constraints:

$$0 \leqslant \varepsilon \leqslant 1$$

$$(7.22) \quad d^T \nabla c_i(x) + \tau_i c_i(x) \begin{cases} = 0 \text{ for } i = 1,2,\ldots,m' \\ \geqslant 0 \text{ for } i = m'+1,m'+2,\ldots,m \end{cases}$$

where

$$(7.23) \quad \tau_i = \begin{cases} 1 \text{ for } c_i(x) > 0 \text{ and } i \in \{m'+1,m'+2,\ldots,m\} \\ \varepsilon \text{ otherwise} \end{cases}$$

This step is a necessary preparation to the next, where constraints (7.21) are linearized : it may occur that the linearized constraints are inconsistent, while the original constraints are not. If the solution obtained is $d = 0$, $\varepsilon = 0$, it is concluded that (7.21) is inconsistent.

(ii) Calculate the gradient $\nabla F(x)$. Solve the following quadratic programming problem in $d \in \mathbb{R}_n$: minimize

$$(7.24) \quad d^T \nabla F(x) + \tfrac{1}{2} d^T Bd$$

subject to the constraints (7.22). This step produces a direction d to be followed in search of a solution for the problem of Definition 7.12. At the same time, the Lagrange multipliers corresponding with the constraints (7.22) can be obtained. These m values will be denoted by the m-vector λ.

(iii) Use this λ and m-vector μ, obtained in the previous iteration, to update μ; in obvious notation

$$(7.25) \quad \mu_i \to \max\{|\lambda_i|, (\mu_i + |\lambda_i|)/2\}, \ i = 1,2,\ldots,m$$

Solve the following nonlinear programming problem in $\alpha \in [0,1]$: minimize

$$(7.26) \quad G(\alpha) := F(x+\alpha d) + \mu^T h(x+\alpha d)$$

where $h := (h_1, h_2, \ldots, h_m)$ and $h_i : \mathbb{R}_n \to \mathbb{R}$ is defined by

$$h_i(y) := \begin{cases} |c_i(y)| & \text{for } i = 1,2,\ldots,m' \\ |\min\{0, c_i(y)\}| & \text{for } i = m'+1, m'+2, \ldots, m \end{cases}$$

This step calculates what distance α to go in direction d. The term $\mu^T h(x+\alpha d)$ in (7.26) is the penalty for trespassing constraints (7.21); its specific form was chosen by Powell as a compromise: approximate solutions $x + \alpha d$ which do not (completely) satisfy (7.21) should be allowed, but penalized. This compromise increases the rate of convergence; however, global convergence no longer is guaranteed. Consequently, a test on cycling has to be included: consider the values of μ and of $G(\alpha)$, obtained in any five successive iterations; if μ is constant and G non-decreasing it is concluded that cycling occurs.

(iv) Define the Lagrange function $L : \mathbb{R}_{n+m} \to \mathbb{R}$ by

$$L(x,\lambda) := F(x) - \lambda^T c(x)$$

Calculate

$$\delta = \alpha d$$

$$\gamma = \nabla_x L(x+\delta, \lambda) - \nabla_x L(x, \lambda)$$

$$(7.27) \quad \theta = \begin{cases} 1 & \text{for } \delta^T \gamma \geq 0.2\delta^T B\delta \\ 0.8\delta^T B\delta / (\delta^T B\delta - \delta^T \gamma) & \text{for } \delta^T \gamma < 0.2\delta^T B\delta \end{cases}$$

$$\eta = \theta\gamma + (1-\theta)B\delta$$

Update x and B as follows:

$$(7.28) \quad x \to x + \delta$$

$$(7.29) \qquad B \to B - \frac{B\delta\delta^T B}{\delta^T B\delta} + \frac{\eta\eta^T}{\delta^T \eta}$$

The updating of x is obvious. Formula (7.29) is a generalization of the well-known updating formula of Broyden, Fletcher, Goldfarb and Shanno in the case of unconstrained minimization, which uses γ in stead of η. The great advantage of (7.29) is that it maintains positive definiteness of B (in general I is chosen as initial value for B). The numerical values in (7.27) are based on (limited) practical experience and may be improved.

This general algorithm was used to solve the specific problem of Theorem 7.3, however with the constraints (7.3) replaced by (7.5). The object function F now reads

$$(5.15) \qquad r_\tau(w) = \mu_2 - \sum_{i=0}^{n} \binom{n}{i} m_{i1}^2/m_{i0}$$

It is a function of $w := (x_1,x_2,\ldots,x_m,y_1,y_2,\ldots,y_{m+1})$, since μ_2 and m_{ij}, defined in (5.11)-(5.12), can be written as

$$(7.30) \qquad \mu_2 = \sum_{k=1}^{m} y_k[x_k^2 + (1-x_k)^2] + y_{m+1}/4$$

$$(7.31) \qquad m_{ij} = \sum_{k=1}^{m} y_k[x_k^{i+j}(1-x_k)^{n-i} + x_k^{n-i}(1-x_k)^{i+j}] + y_{m+1}/2^{n+j},$$

for $i = 0,1,\ldots,n$ and $j = 0$ or 1.

The first partial derivatives of object function and constraints are easy to derive. The simplest expression for the derivative of $r_\tau(w)$ with respect to any variable z can be found by use of the symmetry properties of the prior distributions; they imply for $i = 0,1,\ldots,n$

$$m_{n-i,0} = m_{i0}, \quad m_{n-i,1} = m_{i0} - m_{i1}, \quad d_\tau(n-i) = 1 - d_\tau(i)$$

where the last equality follows from (5.14). The result is

$$(7.32) \qquad \frac{\partial r_\tau(w)}{\partial z} = [1 - 2d_\tau^2(i)]\,\frac{\partial m_{i0}}{\partial z} + 2[2d_\tau(i) - 1]\,\frac{\partial m_{i1}}{\partial z}$$

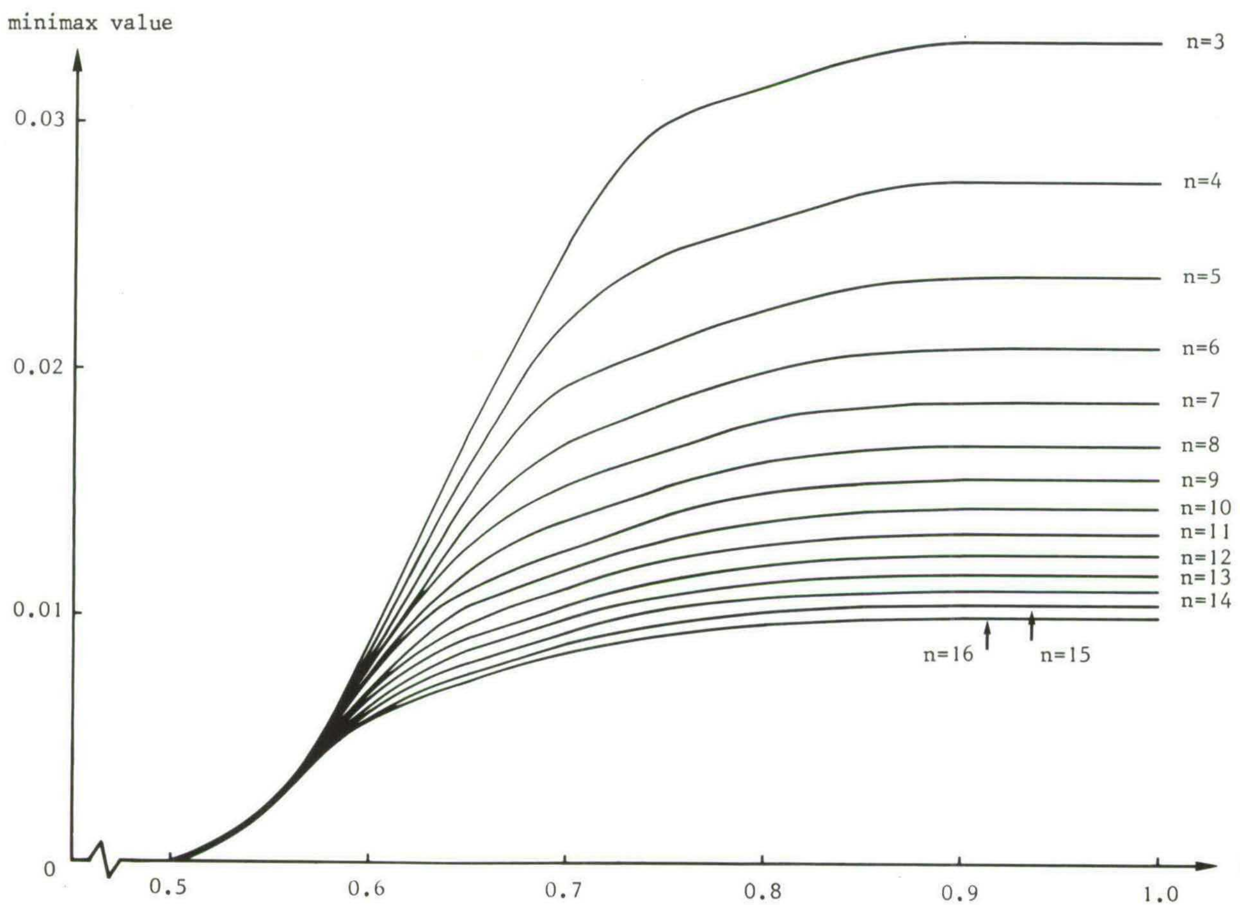
The detailed results for n=3 up to 16 are presented in Tables A1-A14 in Appendix A; they are summarized in Figures 7.13-7.17, which will be briefly discussed now.

The behavior of the minimax value $\max_\tau r_\tau$ as a function of P is shown in Figure 7.13; since it is based on the data in Appendix A, the curve is somewhat unreliable between the successive values of P occurring in the tables. Note that the minimax value is practically constant for $0.9 \leqslant P \leqslant 1$.

Figures 7.14 and 7.15 show some of the minimax estimators d_m for $i \leqslant n/2$. To give a clearer picture, the values calculated in the Appendix are connected by line segments. For small P-values some of the estimates coincide as indicated in the graphs. Note that again $d_m(i)$ is sometimes increasing in P, even for i as small as 1; in general, the behavior of $d_m(i)$ is rather irregular.

For the case P=1, Figure 7.16 presents in detail the least favorable symmetric prior distribution with as few mass points as possible; point masses in $\frac{1}{2}$ are indicated differently. Of course, the connecting line segments have no meaning, apart from increasing readibility. For general P, Figure 7.17 outlines how the minimum number of mass points of these distributions varies with increasing n and P.

Minimax value and minimax rule for P=1 were analytically obtained in Example 1.14. Comparison of these exact values with the numerical results in Appendix A gives an indication of the numerical precision of the calculations. Define δ_r as the absolute difference between the calculated value and the theoretical value of $\max_\tau r_\tau$; define δ_d as the maximum (with respect to i) of the absolute difference between the calculated value and the theoretical value of $d_m(i)$. Table 7.18 shows some values of these indicators of numerical accuracy.



<u>Figure 7.13</u>. Minimax value $\max_{\tau} r_{\tau}$ as function of P for different n

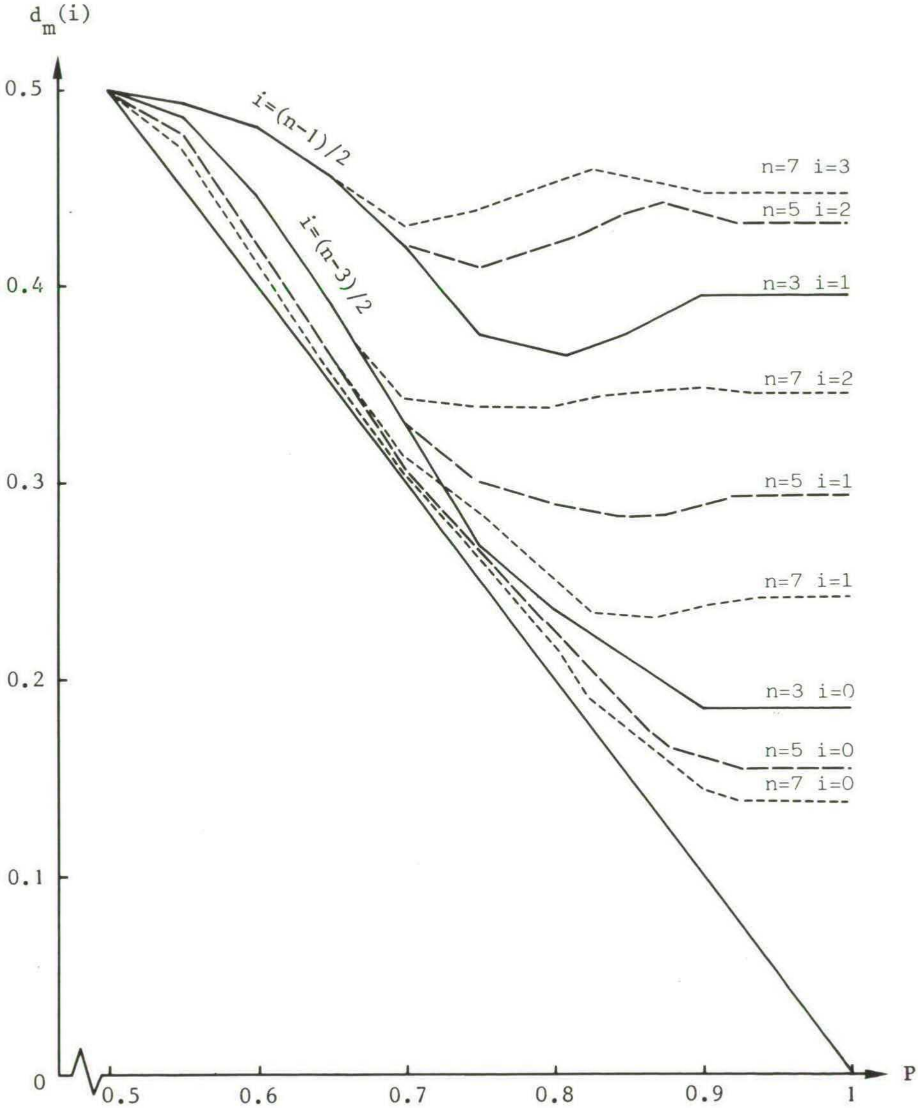


Figure 7.14. Minimax estimators d_m for odd n

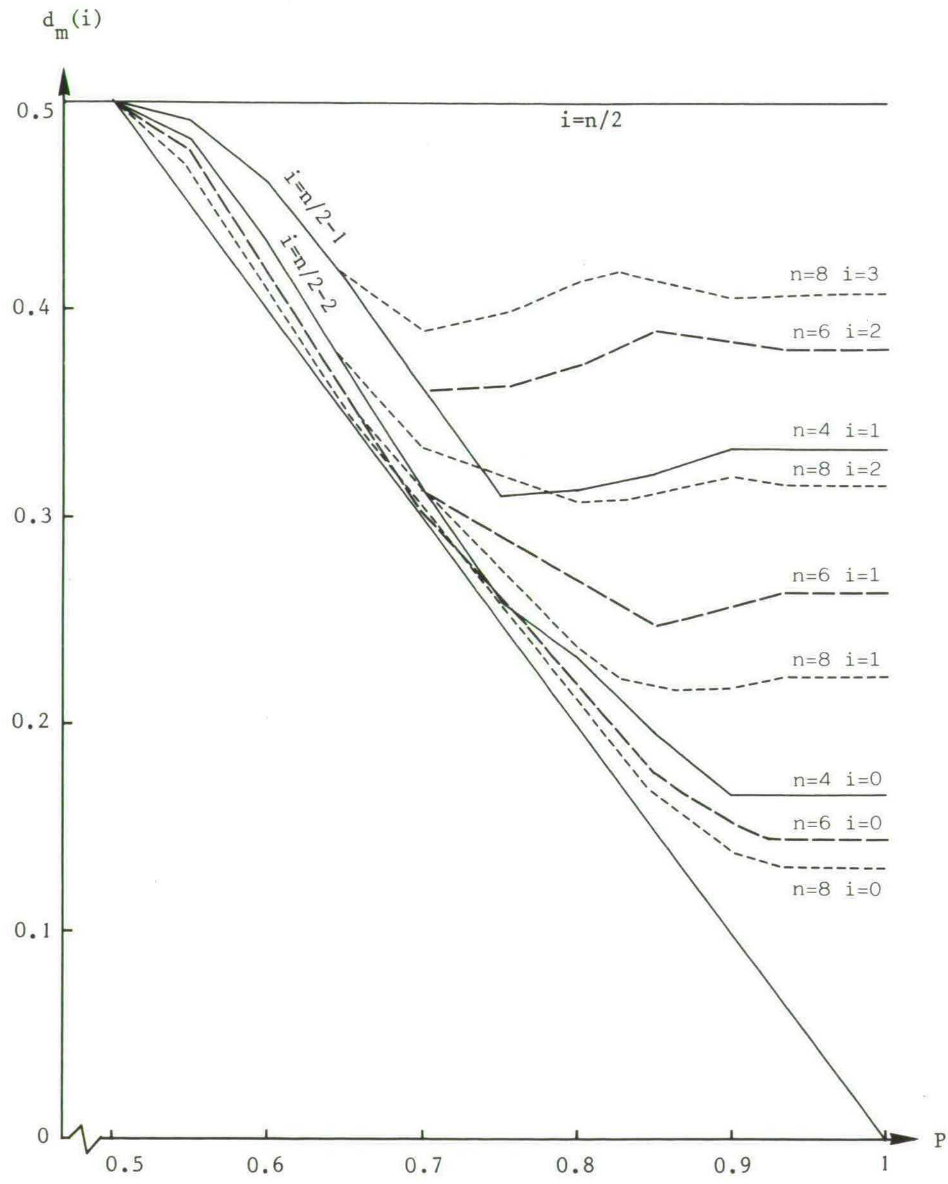


Figure 7.15. Minimax estimators d_m for even n

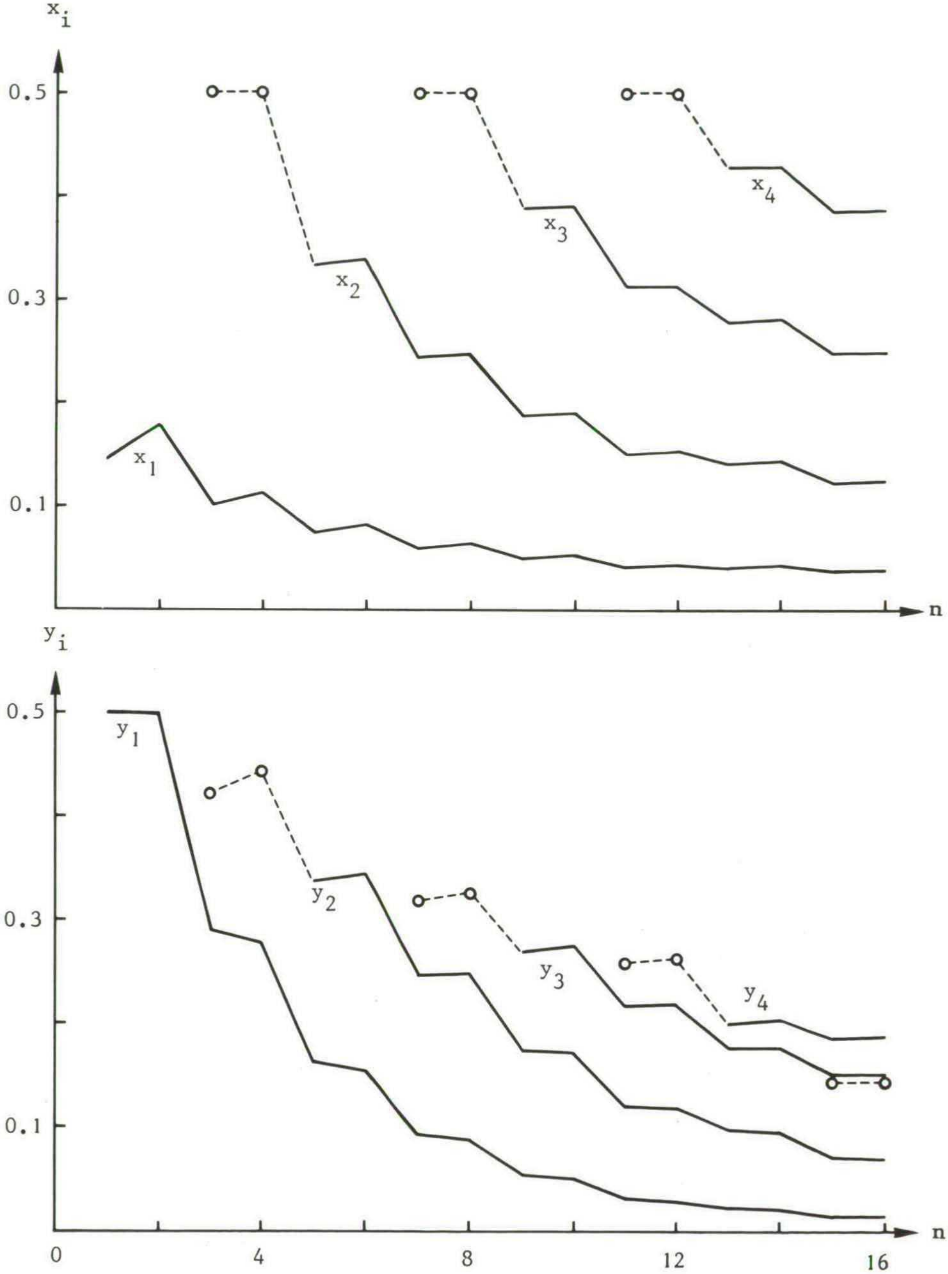


Figure 7.16. Least favorable symmetric prior distributions for P=1 and different n : locations x_i ($< \frac{1}{2}$) and point masses y_i

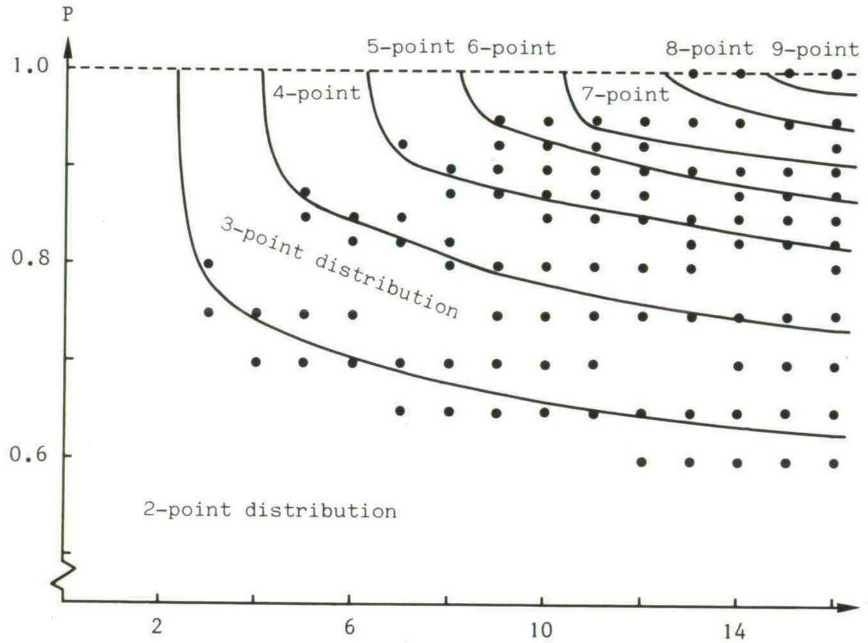


Figure 7.17. Minimum number of steps of least favorable symmetric prior distributions

Table 7.18. Numerical accuracy of minimax calculations

n	$10^{11}*\delta_r$	$10^7*\delta_d$
12	1	0
13	31	235
14	0	565
15	2	81
16	17	193

7.4. Weighted quadratic loss

This final section discusses the truncated binomial problem of Definition 5.1 with the following slight modification, however: instead of the usual quadratic loss $(\theta-a)^2$, the weighted quadratic loss function

$$(7.33) \qquad L(\theta,a) = \frac{(\theta-a)^2}{\theta(1-\theta)}$$

will be used now. As was mentioned previously, an interesting consequence is that in the nontruncated case the sample fraction X/n is a minimax estimator for θ with (constant) risk $1/n$.

In the truncated situation the following counterpart of Theorem 5.11 is easily derived using the definitions

$$(7.34) \qquad k_{ij} := E_\tau[\theta^{i+j-1}(1-\theta)^{n-i}] \ , \ i = 0,1,\ldots,n \ ; \ j = 0,1,2,\ldots$$

Theorem 7.19. For the truncated binomial problem with, however, the loss function (7.33), the nonrandomized Bayes rule with respect to some $\tau \in \Theta^*$ is given by

$$(7.35) \qquad d_\tau(i) = k_{i1}/k_{i0} \ , \ i = 0,1,\ldots,n$$

Its (minimum) Bayes risk equals

$$(7.36) \qquad r_\tau := r(\tau,d_\tau) = E_\tau[\theta/(1-\theta)] - \sum_{i=0}^{n} \binom{n}{i} k_{i1}^2/k_{i0}$$

and d_τ is admissible. □

Since the problem is invariant again, all attention may be directed towards symmetric prior distributions, leading to the relations

$$(7.37) \qquad k_{n-1,0} = k_{i0},\ k_{n-i,1} = k_{i0} - k_{i1},\ d_\tau(n-i) = 1 - d_\tau(i)$$

To start with, minimax estimators are considered for n=1 up to 3.

Example 7.20. For n=1, by means of (7.34) and (7.37), (7.36) reduces to

$$r_\tau = 1 - 2/\upsilon$$

$$\upsilon := E_\tau(1/\theta) \tag{7.38}$$

For symmetric priors the maximum of υ equals $2/(1-\phi)$ and is attained for the distribution (7.10). Consequently, the minimax value is ϕ and minimax estimator d_m is determined by $d_m(1) = (\phi+1)/2$. The same result can be obtained by the method used in Example 2.12.

For n=2, (7.36) becomes after some simplifications

$$r_\tau = [1 - 1/(\upsilon-1)]/2$$

so that (7.10) is least favorable again. The minimax value is $\phi/(1+\phi)$ and $d_m(1) = \frac{1}{2}(3\phi+1)/(\phi+1)$. Note that both minimax estimators are identical to the ones presented in Table 7.9 for 'small' values of P.

For n=3, some manipulation reduces (7.36) to

$$r_\tau = [(1+3m_2)(1-m_2) - \frac{(1+m_2)^2}{2\upsilon-3}]/4$$

which is increasing in υ, but decreasing in m_2. With some effort it can be derived that (7.10) is least favorable, provided that

$$13\phi^3 + 5\phi^2 - \phi - 1 \leqslant 0$$

holds, or, equivalently, $\phi \leqslant 0.373$. Further, it can be seen that for P=1 the distribution

$$\begin{Bmatrix} 0 \\ 1/6 \end{Bmatrix}_S$$

is least favorable. □

In general, r_τ depends on τ only through the vector

$$\mu^* := (\upsilon,\mu_0,\mu_1,\ldots,\mu_n)$$

This leads to the analysis of the Tchebycheff system of functions f_i : $[1-P,P] \to \mathbb{R}$, defined by

$$f_i(u) = u^i \ , \ i = -1,0,1,\ldots,n$$

Now generalize moment space D_B^r of Definition 6.1 as follows: D_B^r consists of all vectors $(c_{-1},c_0,c_1,\ldots,c_r)$ satisfying

$$c_i = \int_B u^i dF(u) \ , \ i = -1,0,1,\ldots,r$$

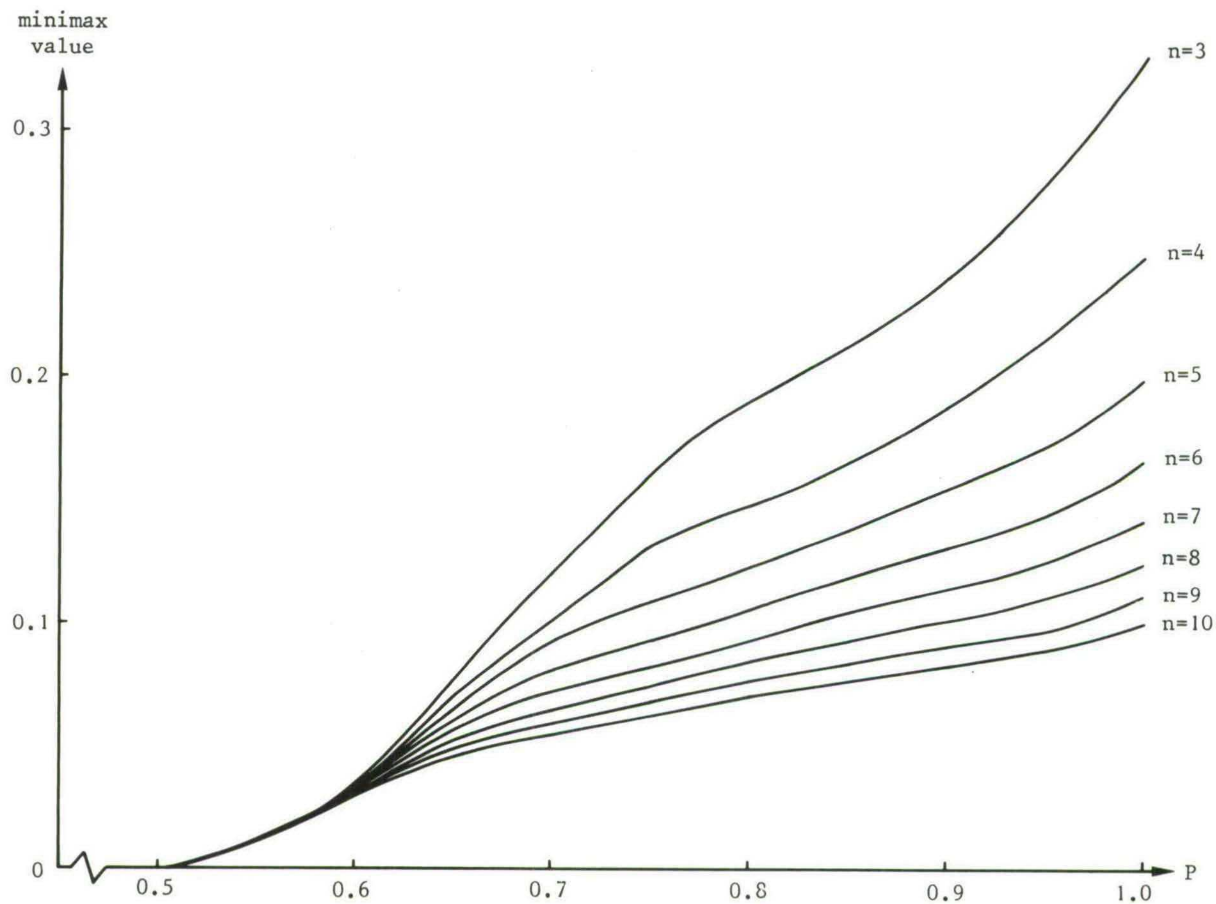
Then the general analysis of KARLIN & STUDDEN 1966 shows that Theorems 6.5 and 6.7 remain valid; the same holds for Theorem 6.12. Therefore the computer program, described in Section 7.3, could easily be adapted to the present situation, the main modification being the replacement of the object function (5.15) by (7.36).

Detailed results can be found in Appendix B for n=3 up to 10. Pictures, similar to Figures 7.13-7.17 can be drawn, but only the behavior of the minimax value is shown here (Figure 7.21). For the rest, some interesting features will be noted below, particularly stressing the similarities and differences with the previous case of unweighted quadratic loss.

(i) The necessary number of mass points for a symmetric least favorable prior distribution gradually increases with P. The maximum number of locations smaller that $\frac{1}{2}$ satisfies (7.6), as in the previous case. Unlike the previous case, a point mass occurs at 1-P for all values of n and P.

(ii) For P=1 the least favorable priors for n and n+1 (n odd), respectively, are identical.

(iii) The behavior of the minimax estimates $d_m(i)$ as function of P is again very irregular. For small values of P, the minimax rules corresponding with weighted and unweighted quadratic loss respectively are identical.



<u>Figure 7.21</u>. Minimax value $\max_\tau r_\tau$ for weighted quadratic loss function (7.33)

Appendix A

This appendix contains fourteen tables, presenting minimax estimators for the following estimation problem (Θ,L,X): parameter space Θ equals $[1-P,P]$ with $\frac{1}{2} < P \leq 1$, $X \in B(n,\theta)$ and the loss function is given by

$$L(\theta,a) = (\theta-a)^2$$

(This is the symmetric truncated binomial problem of Definition 5.1, where h is the identity.)

Each table successively presents for a certain n ($3 \leq n \leq 16$) and several values of P the following elements:

(i) the (symmetric) minimax estimator d_m, determined by the estimates $d_m(i)$ for $i = 0,1,\ldots,\mathrm{ent}(n/2)$;

(ii) the minimax value $\max_\tau r_\tau$;

(iii) the (discrete) symmetric least favorable prior distribution with the minimum number of steps - determined by (7.6).

<u>Tabel A1</u>. Case n = 3: minimax solution for quadratic loss

P	minimax rule d_m		minimax	symmetric least
	$d_m(0)$	$d_m(1)$	value (* 100)	favorable prior
0.6	0.44571	0.48000	0.88869	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.39051	0.45500	1.73119	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.32919	0.42000	2.51728	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.26786	0.37500	3.01339	$\begin{Bmatrix}0.25\\0.5\end{Bmatrix}_S$
0.8	0.23472	0.36284	3.17026	$\begin{Bmatrix}0.2\\0.4167\end{Bmatrix}_S$
0.825	0.22306	0.36749	3.23439	$\begin{Bmatrix}0.175\\0.3723\end{Bmatrix}_S$
0.85	0.21024	0.37391	3.29148	$\begin{Bmatrix}0.15\\0.3376\end{Bmatrix}_S$
0.875	0.19640	0.38284	3.33374	$\begin{Bmatrix}0.125\\0.3100\end{Bmatrix}_S$
0.9	0.18301	0.39434	3.34936	$\begin{Bmatrix}0.1019\\0.2887\end{Bmatrix}_S$
1	0.18301	0.39434	3.34936	$\begin{Bmatrix}0.1019\\0.2887\end{Bmatrix}_S$

Table A2. Case n = 4: minimax solution for quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	minimax value (* 100)	symmetric least favorable prior
0.6	0.43299	0.46154	0.85646	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.37326	0.41743	1.60106	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.31305	0.36207	2.20567	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.26166	0.31200	2.47725	$\begin{Bmatrix}0.25\\0.4718\end{Bmatrix}_S$
0.8	0.23250	0.31440	2.61689	$\begin{Bmatrix}0.2\\0.3647\end{Bmatrix}_S$
0.825	0.21560	0.31716	2.68263	$\begin{Bmatrix}0.175\\0.3305\end{Bmatrix}_S$
0.85	0.19712	0.32161	2.73785	$\begin{Bmatrix}0.15\\0.3047\end{Bmatrix}_S$
0.875	0.17708	0.32858	2.77264	$\begin{Bmatrix}0.125\\0.2854\end{Bmatrix}_S$
0.9	0.16667	0.33333	2.77778	$\begin{Bmatrix}0.1127\\0.2778\end{Bmatrix}_S$
1	0.16667	0.33333	2.77778	$\begin{Bmatrix}0.1127\\0.2778\end{Bmatrix}_S$

Table A3. Case n = 5: minimax solution for quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	minimax value (* 100)	symmetric least favorable prior
0.6	0.42327	0.44571	0.48000	0.82614	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.36299	0.39051	0.45500	1.48459	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.30570	0.32919	0.42000	1.94059	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.26503	0.29996	0.40906	2.11246	$\begin{Bmatrix}0.25\\0.4130\end{Bmatrix}_S$
0.8	0.22695	0.28939	0.41954	2.25669	$\begin{Bmatrix}0.2\\0.3318\end{Bmatrix}_S$
0.825	0.20574	0.28507	0.42712	2.31680	$\begin{Bmatrix}0.175\\0.3062\end{Bmatrix}_S$
0.85	0.18330	0.28235	0.43653	2.53900	$\begin{Bmatrix}0.15\\0.2867\end{Bmatrix}_S$
0.875	0.16690	0.28356	0.44088	2.37627	$\begin{Bmatrix}0.125 & 0.4191\\0.2533 & 0.2467\end{Bmatrix}_S$
0.9	0.15971	0.28780	0.43653	2.38452	$\begin{Bmatrix}0.1 & 0.3644\\0.2040 & 0.2960\end{Bmatrix}_S$
0.925	0.15451	0.29271	0.43090	2.38729	$\begin{Bmatrix}0.0760 & 0.3341\\0.1633 & 0.3367\end{Bmatrix}_S$
1	0.15451	0.29271	0.43090	2.38729	$\begin{Bmatrix}0.0760 & 0.3341\\0.1633 & 0.3367\end{Bmatrix}_S$

Table A4. Case n = 6: minimax solution for quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	minimax value (* 100)	symmetric least favorable prior
0.6	0.41614	0.43299	0.46154	0.79749	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.35714	0.37326	0.41743	1.37959	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.30246	0.31305	0.36207	1.71519	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.26421	0.29157	0.36350	1.86168	$\begin{Bmatrix}0.25\\0.3768\end{Bmatrix}_S$
0.8	0.22037	0.26892	0.37329	1.99812	$\begin{Bmatrix}0.2\\0.3112\end{Bmatrix}_S$
0.825	0.19661	0.25738	0.38169	2.04743	$\begin{Bmatrix}0.175\\0.2913\end{Bmatrix}_S$
0.85	0.17558	0.25002	0.38949	2.07325	$\begin{Bmatrix}0.15 & 0.4416\\0.2654 & 0.2346\end{Bmatrix}_S$
0.875	0.16387	0.25292	0.38799	2.08836	$\begin{Bmatrix}0.125 & 0.3838\\0.2180 & 0.2820\end{Bmatrix}_S$
0.9	0.15261	0.25793	0.38509	2.09841	$\begin{Bmatrix}0.1 & 0.3537\\0.1792 & 0.3208\end{Bmatrix}_S$
0.925	0.14495	0.26330	0.38165	2.10102	$\begin{Bmatrix}0.0817 & 0.3375\\0.1551 & 0.3449\end{Bmatrix}_S$
1	0.14495	0.26330	0.38165	2.10102	$\begin{Bmatrix}0.0817 & 0.3375\\0.1551 & 0.3449\end{Bmatrix}_S$

Table A5. Case n = 7: minimax solution for quadratic loss

P	minimax rule d_m				minimax	symmetric least
	$d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	value (* 100)	favorable prior
0.65	0.35389	0.36299	0.39051	0.45500	1.28429	$\begin{Bmatrix} 0.35 \\ 0.5 \end{Bmatrix}_S$
0.7	0.30421	0.31265	0.34205	0.43004	1.53003	$\begin{Bmatrix} 0.3 \\ 0.4608 \end{Bmatrix}_S$
0.75	0.26203	0.28406	0.33773	0.43765	1.67514	$\begin{Bmatrix} 0.25 \\ 0.3512 \end{Bmatrix}_S$
0.8	0.21453	0.25108	0.33843	0.44998	1.79845	$\begin{Bmatrix} 0.2 \\ 0.2974 \end{Bmatrix}_S$
0.825	0.18949	0.23378	0.34266	0.45801	1.83460	$\begin{Bmatrix} 0.175 \\ 0.2815 \end{Bmatrix}_S$
0.85	0.17412	0.23139	0.34487	0.45559	1.85431	$\begin{Bmatrix} 0.15 & 0.4049 \\ 0.2349 & 0.2651 \end{Bmatrix}_S$
0.875	0.15953	0.23225	0.34656	0.45171	1.86964	$\begin{Bmatrix} 0.125 & 0.3716 \\ 0.1950 & 0.3050 \end{Bmatrix}_S$
0.9	0.14487	0.23609	0.34747	0.44693	1.87892	$\begin{Bmatrix} 0.1 & 0.3499 \\ 0.1631 & 0.3369 \end{Bmatrix}_S$
0.925	0.13802	0.23979	0.34582	0.44669	1.88075	$\begin{Bmatrix} 0.075 & 0.2764 \\ 0.1197 & 0.2524 \end{Bmatrix}_S$
0.95	0.13715	0.24082	0.34449	0.44816	1.88090	$\begin{Bmatrix} 0.0601 & 0.2437 \\ 0.0931 & 0.2476 \end{Bmatrix}_S$
1	0.13715	0.24082	0.34449	0.44816	1.88090	$\begin{Bmatrix} 0.0601 & 0.2437 \\ 0.0931 & 0.2476 \end{Bmatrix}_S$

Table A6. Case n = 8: minimax solution for quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	minimax value (* 100)	symmetric least favorable prior
0.65	0.35211	0.35714	0.37326	0.41743	1.19741	$\begin{Bmatrix} 0.35 \\ 0.5 \end{Bmatrix}_S$
0.7	0.30499	0.31258	0.33348	0.38965	1.38862	$\begin{Bmatrix} 0.3 \\ 0.4267 \end{Bmatrix}_S$
0.75	0.25952	0.27693	0.31922	0.39766	1.52930	$\begin{Bmatrix} 0.25 \\ 0.3326 \end{Bmatrix}_S$
0.8	0.20996	0.23633	0.30757	0.41368	1.63626	$\begin{Bmatrix} 0.2 \\ 0.2879 \end{Bmatrix}_S$
0.825	0.18913	0.22339	0.30764	0.41742	1.66220	$\begin{Bmatrix} 0.175 & 0.4298 \\ 0.2541 & 0.2459 \end{Bmatrix}_S$
0.85	0.17197	0.21851	0.31132	0.41446	1.68188	$\begin{Bmatrix} 0.15 & 0.3899 \\ 0.2123 & 0.2877 \end{Bmatrix}_S$
0.875	0.15464	0.21618	0.31588	0.41011	1.69643	$\begin{Bmatrix} 0.125 & 0.3661 \\ 0.1788 & 0.3212 \end{Bmatrix}_S$
0.9	0.13858	0.21751	0.31972	0.40538	1.70347	$\begin{Bmatrix} 0.1 & 0.3334 \\ 0.1483 & 0.3029 \end{Bmatrix}_S$
0.925	0.13236	0.22124	0.31692	0.40668	1.70549	$\begin{Bmatrix} 0.075 & 0.2656 \\ 0.1048 & 0.2480 \end{Bmatrix}_S$
0.95	0.13060	0.22295	0.31530	0.40765	1.70569	$\begin{Bmatrix} 0.0637 & 0.2470 \\ 0.0878 & 0.2490 \end{Bmatrix}_S$
1	0.13060	0.22295	0.31530	0.40765	1.70569	$\begin{Bmatrix} 0.0637 & 0.2470 \\ 0.0878 & 0.2490 \end{Bmatrix}_S$

<u>Table A7</u>. Case n = 9: minimax solution for quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	minimax value (* 100)	least favorable symmetric prior
0.65	0.35114	0.35389	0.36299	0.39051	0.45500	1.11785	$\left\{\begin{matrix} 0.35 \\ 0.5 \end{matrix}\right\}_S$
0.7	0.30490	0.31163	0.32797	0.36697	0.44668	1.27643	$\left\{\begin{matrix} 0.3 \\ 0.4003 \end{matrix}\right\}_S$
0.75	0.25721	0.27058	0.30402	0.36886	0.45531	1.41076	$\left\{\begin{matrix} 0.25 \\ 0.3187 \end{matrix}\right\}_S$
0.8	0.20729	0.22619	0.28127	0.37943	0.46718	1.49983	$\left\{\begin{matrix} 0.2 & 0.4714 \\ 0.2775 & 0.2225 \end{matrix}\right\}_S$
0.825	0.18841	0.21635	0.28277	0.37977	0.46451	1.52303	$\left\{\begin{matrix} 0.175 & 0.4100 \\ 0.2318 & 0.2682 \end{matrix}\right\}_S$
0.85	0.16926	0.20813	0.28644	0.37921	0.46107	1.54195	$\left\{\begin{matrix} 0.15 & 0.3824 \\ 0.1956 & 0.3044 \end{matrix}\right\}_S$
0.875	0.14964	0.20218	0.29234	0.37733	0.45681	1.55479	$\left\{\begin{matrix} 0.125 & 0.3636 \\ 0.1671 & 0.3329 \end{matrix}\right\}_S$
0.9	0.13565	0.20241	0.29488	0.37525	0.45615	1.56013	$\left\{\begin{matrix} 0.1 & 0.3061 \\ 0.1308 & 0.2588 \end{matrix}\right\}_S$
0.925	0.12709	0.20655	0.29307	0.37436	0.45841	1.56233	$\left\{\begin{matrix} 0.075 & 0.2606 \\ 0.0937 & 0.2482 \end{matrix}\right\}_S$
0.95	0.12500	0.20833	0.29167	0.37499	0.45835	1.56250	$\left\{\begin{matrix} 0.0500 & 0.1893 & 0.3893 \\ 0.0545 & 0.1752 & 0.2704 \end{matrix}\right\}_S$
1	0.12500	0.20833	0.29167	0.37500	0.45833	1.56520	$\left\{\begin{matrix} 0.0498 & 0.1888 & 0.3890 \\ 0.0542 & 0.1749 & 0.2709 \end{matrix}\right\}_S$

Table A8. Case n = 10: minimax solution for quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	minimax value (* 100)	least favorable symmetric prior
0.65	0.35061	0.35211	0.35714	0.37326	0.41743	1.04479	$\left\{\begin{matrix}0.35\\0.5\end{matrix}\right\}_S$
0.7	0.30438	0.31017	0.32343	0.35277	0.41099	1.18488	$\left\{\begin{matrix}0.3\\0.3795\end{matrix}\right\}_S$
0.75	0.25529	0.26528	0.29121	0.34523	0.42231	1.31144	$\left\{\begin{matrix}0.25\\0.3081\end{matrix}\right\}_S$
0.8	0.20746	0.22311	0.26535	0.34599	0.43404	1.38549	$\left\{\begin{matrix}0.2 & 0.4357\\0.2552 & 0.2448\end{matrix}\right\}_S$
0.825	0.18714	0.21047	0.26467	0.34854	0.43074	1.40790	$\left\{\begin{matrix}0.175 & 0.3998\\0.2145 & 0.2855\end{matrix}\right\}_S$
0.85	0.16632	0.19900	0.26684	0.35099	0.42619	1.42552	$\left\{\begin{matrix}0.15 & 0.3783\\0.1830 & 0.3170\end{matrix}\right\}_S$
0.875	0.14553	0.18994	0.27208	0.35241	0.42091	1.43596	$\left\{\begin{matrix}0.125 & 0.3558\\0.1568 & 0.3186\end{matrix}\right\}_S$
0.9	0.13270	0.19053	0.27382	0.34961	0.42228	1.44082	$\left\{\begin{matrix}0.1 & 0.2932\\0.1174 & 0.2486\end{matrix}\right\}_S$
0.925	0.12188	0.19447	0.27373	0.34673	0.42478	1.44286	$\left\{\begin{matrix}0.075 & 0.2586\\0.0853 & 0.2496\end{matrix}\right\}_S$
0.95	0.12013	0.19610	0.27208	0.34805	0.42403	1.44304	$\left\{\begin{matrix}0.0520 & 0.1910 & 0.3899\\0.0505 & 0.1733 & 0.2762\end{matrix}\right\}_S$
1	0.12013	0.19610	0.27208	0.34805	0.42403	1.44304	$\left\{\begin{matrix}0.0520 & 0.1910 & 0.3899\\0.0505 & 0.1733 & 0.2762\end{matrix}\right\}_S$

Table A9. Case n = 11: minimax solution for quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	$d_m(5)$	minimax value (* 100)	least favorable symmetric prior
0.65	0.35033	0.35114	0.35389	0.36299	0.39051	0.45500	0.97749	$\begin{Bmatrix} 0.35 \\ 0.5 \end{Bmatrix}_S$
0.7	0.30370	0.30853	0.31937	0.34252	0.38711	0.45798	1.10837	$\begin{Bmatrix} 0.3 \\ 0.3628 \end{Bmatrix}_S$
0.75	0.25380	0.26108	0.28067	0.32475	0.39472	0.46721	1.22619	$\begin{Bmatrix} 0.25 \\ 0.2999 \end{Bmatrix}_S$
0.8	0.20716	0.22040	0.25409	0.31991	0.40206	0.47080	1.28935	$\begin{Bmatrix} 0.2 & 0.4206 \\ 0.2371 & 0.2629 \end{Bmatrix}_S$
0.825	0.18560	0.20520	0.25054	0.32361	0.40030	0.46778	1.31071	$\begin{Bmatrix} 0.175 & 0.3940 \\ 0.2010 & 0.2990 \end{Bmatrix}_S$
0.85	0.16346	0.19072	0.25016	0.32838	0.39716	0.46405	1.32657	$\begin{Bmatrix} 0.15 & 0.3761 \\ 0.1734 & 0.3266 \end{Bmatrix}_S$
0.875	0.14421	0.18185	0.25314	0.33006	0.39455	0.46265	1.33503	$\begin{Bmatrix} 0.125 & 0.3314 \\ 0.1421 & 0.2668 \end{Bmatrix}_S$
0.9	0.12967	0.18073	0.25629	0.32715	0.39435	0.46455	1.33973	$\begin{Bmatrix} 0.1 & 0.2864 \\ 0.1070 & 0.2462 \end{Bmatrix}_S$
0.925	0.11865	0.18340	0.25736	0.32452	0.39510	0.46587	1.34147	$\begin{Bmatrix} 0.075 & 0.2408 & 0.4255 \\ 0.0752 & 0.2096 & 0.2152 \end{Bmatrix}_S$
0.95	0.11584	0.18567	0.25556	0.32531	0.39534	0.46493	1.34169	$\begin{Bmatrix} 0.0515 & 0.1814 & 0.3534 \\ 0.0436 & 0.1527 & 0.2180 \end{Bmatrix}_S$
1	0.11583	0.18568	0.25553	0.32538	0.39523	0.46508	1.34169	$\begin{Bmatrix} 0.0419 & 0.1514 & 0.3119 \\ 0.0315 & 0.1204 & 0.2180 \end{Bmatrix}_S$

Table A10. Case n = 12: minimax solution for quadratic loss

P	minimax rule d_m: $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	$d_m(5)$	minimax value (* 100)	least favorable symmetric prior
0.65	0.35048	0.35116	0.35311	0.35890	0.37598	0.42016	0.91582	$\begin{Bmatrix} 0.35 \\ 0.4887 \end{Bmatrix}_S$
0.7	0.30302	0.30694	0.31572	0.33432	0.36989	0.42702	1.04318	$\begin{Bmatrix} 0.3 \\ 0.3492 \end{Bmatrix}_S$
0.75	0.25269	0.25789	0.27232	0.30714	0.36947	0.44073	1.15159	$\begin{Bmatrix} 0.25 \\ 0.2934 \end{Bmatrix}_S$
0.8	0.20655	0.21781	0.24545	0.29991	0.37367	0.44238	1.20720	$\begin{Bmatrix} 0.2 & 0.4119 \\ 0.2225 & 0.2775 \end{Bmatrix}_S$
0.825	0.18399	0.20036	0.23874	0.30337	0.37449	0.43803	1.22726	$\begin{Bmatrix} 0.175 & 0.3904 \\ 0.1903 & 0.3097 \end{Bmatrix}_S$
0.85	0.16083	0.18327	0.23517	0.30923	0.37421	0.43250	1.24100	$\begin{Bmatrix} 0.15 & 0.3750 \\ 0.1658 & 0.3342 \end{Bmatrix}_S$
0.875	0.14279	0.17529	0.23757	0.30979	0.37195	0.43302	1.24841	$\begin{Bmatrix} 0.125 & 0.3180 \\ 0.1302 & 0.2512 \end{Bmatrix}_S$
0.9	0.12658	0.17220	0.24169	0.30774	0.36987	0.43535	1.25281	$\begin{Bmatrix} 0.1 & 0.2827 \\ 0.0988 & 0.2464 \end{Bmatrix}_S$
0.925	0.11569	0.17406	0.24299	0.30571	0.37009	0.43599	1.25427	$\begin{Bmatrix} 0.075 & 0.2304 & 0.4118 \\ 0.0671 & 0.1911 & 0.2418 \end{Bmatrix}_S$
0.95	0.11204	0.17659	0.24150	0.30574	0.37097	0.43512	1.25450	$\begin{Bmatrix} 0.0539 & 0.1845 & 0.3553 \\ 0.0417 & 0.1517 & 0.2203 \end{Bmatrix}_S$
1	0.11200	0.17667	0.24134	0.30600	0.37067	0.43533	1.25450	$\begin{Bmatrix} 0.0437 & 0.1536 & 0.3134 \\ 0.0295 & 0.1184 & 0.2199 \end{Bmatrix}_S$

Table A11. Case n = 13: minimax solution for quadratic loss

P	$d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	$d_m(5)$	$d_m(6)$	minimax value (* 100)	least favorable symmetric prior
	minimax rule d_m								
0.65	.35084	.35170	.35363	.35833	.37032	.40000	.46026	0.86240	$\begin{Bmatrix} .35 \\ .4650 \end{Bmatrix}_S$
0.7	.30239	.30552	.31252	.32747	.35640	.40375	.46617	0.98669	$\begin{Bmatrix} .3 \\ .3379 \end{Bmatrix}_S$
0.75	.25187	.25554	.26594	.29252	.34593	.41619	.47563	1.08529	$\begin{Bmatrix} .25 \\ .2882 \end{Bmatrix}_S$
0.8	.20578	.21533	.23834	.28415	.34974	.41565	.47302	1.13601	$\begin{Bmatrix} .2 & .4064 \\ .2105 & .2895 \end{Bmatrix}_S$
0.825	.18244	.19598	.22845	.28615	.35299	.41208	.46977	1.15454	$\begin{Bmatrix} .175 & .3882 \\ .1817 & .3183 \end{Bmatrix}_S$
0.85	.15983	.17880	.22284	.29047	.35457	.40838	.46731	1.16613	$\begin{Bmatrix} .15 & .3579 \\ .1549 & .2866 \end{Bmatrix}_S$
0.875	.14126	.16967	.22468	.29186	.35160	.40843	.46884	1.17316	$\begin{Bmatrix} .125 & .3101 \\ .1204 & .2458 \end{Bmatrix}_S$
0.9	.12346	.16440	.22922	.29134	.34813	.40970	.47092	1.17711	$\begin{Bmatrix} .1 & .2808 & .4969 \\ .0923 & .2472 & .1605 \end{Bmatrix}_S$
0.925	.11290	.16608	.23037	.28929	.34885	.40996	.47015	1.17841	$\begin{Bmatrix} .075 & .2242 & .4060 \\ .0606 & .1816 & .2578 \end{Bmatrix}_S$
0.95	.10868	.16864	.22918	.28906	.34955	.40965	.46984	1.17863	$\begin{Bmatrix} .05 & .1658 & .3236 \\ .0323 & .1240 & .2170 \end{Bmatrix}_S$
1	.10856	.16878	.22901	.28922	.34946	.40965	.46991	1.17863	$\begin{Bmatrix} .0417 & .1417 & .2804 & .4288 \\ .0238 & .0979 & .1773 & .2010 \end{Bmatrix}_S$

Table A12. Case n = 14: minimax solution for quadratic loss

P	minimax rule d_m							minimax	least favorable
	$d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	$d_m(5)$	$d_m(6)$	value (* 100)	symmetric prior
0.65	.35099	.35191	.35378	.35779	.36686	.38766	.43119	0.81625	$\begin{Bmatrix} .35 \\ .4449 \end{Bmatrix}_S$
0.7	.30186	.30430	.30981	.32171	.34528	.38521	.43953	0.93706	$\begin{Bmatrix} .3 \\ .3285 \end{Bmatrix}_S$
0.75	.25141	.25408	.26163	.28151	.32504	.39137	.45421	1.02563	$\begin{Bmatrix} .25 & .4833 \\ .2821 & .2179 \end{Bmatrix}_S$
0.8	.20497	.21299	.23223	.27122	.32975	.39184	.44744	1.07355	$\begin{Bmatrix} .2 & .4028 \\ .2006 & .2994 \end{Bmatrix}_S$
0.825	.18104	.19209	.21934	.27083	.33460	.39050	.44239	1.09038	$\begin{Bmatrix} .175 & .3868 \\ .1747 & .3253 \end{Bmatrix}_S$
0.85	.15925	.17570	.21322	.27358	.33592	.38834	.44086	1.10044	$\begin{Bmatrix} .15 & .3432 \\ .1440 & .2599 \end{Bmatrix}_S$
0.875	.13964	.16463	.21376	.27629	.33324	.38677	.44285	1.10708	$\begin{Bmatrix} .125 & .3052 \\ .1223 & .2443 \end{Bmatrix}_S$
0.9	.12158	.15801	.21754	.27711	.33005	.38644	.44491	1.11050	$\begin{Bmatrix} .1 & .2703 & .4452 \\ .0848 & .2225 & .1927 \end{Bmatrix}_S$
0.925	.11017	.15912	.21936	.27472	.33039	.38751	.44370	1.11174	$\begin{Bmatrix} .075 & .2206 & .4032 \\ .0552 & .1764 & .2684 \end{Bmatrix}_S$
0.95	.10558	.16165	.21836	.27438	.33101	.38723	.44364	1.11193	$\begin{Bmatrix} .05 & .1635 & .3207 \\ .0288 & .1186 & .2203 \end{Bmatrix}_S$
1	.10545	.16181	.21819	.27452	.33095	.38721	.44368	1.11194	$\begin{Bmatrix} .0433 & .1437 & .2820 & .4294 \\ .0224 & .0962 & .1781 & .2033 \end{Bmatrix}_S$

<u>Table A13</u>. Case n = 15: minimax solution for quadratic loss

P	minimax rule d_m								minimax	least favorable
	$d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	$d_m(5)$	$d_m(6)$	$d_m(7)$	value (* 100)	symmetric prior
0.65	.35101	.35191	.35366	.35716	.36441	.37979	.41132	.46610	0.77590	$\begin{Bmatrix} .35 \\ .4276 \end{Bmatrix}_S$
0.7	.30143	.30331	.30757	.31692	.33595	.36965	.41774	.47238	0.89292	$\begin{Bmatrix} .3 \\ .3206 \end{Bmatrix}_S$
0.75	.25155	.25402	.26040	.27614	.31035	.36706	.42990	.47947	0.97244	$\begin{Bmatrix} .25 & .4573 \\ .2675 & .2325 \end{Bmatrix}_S$
0.8	.20418	.21085	.22689	.26019	.31273	.37123	.42412	.47447	1.01817	$\begin{Bmatrix} .2 & .4003 \\ .1923 & .3077 \end{Bmatrix}_S$
0.825	.17982	.18872	.21131	.25681	.31803	.37266	.41918	.47110	1.03313	$\begin{Bmatrix} .175 & .3860 \\ .1689 & .3311 \end{Bmatrix}_S$
0.85	.15858	.17296	.20533	.25911	.31863	.37011	.41890	.47201	1.04231	$\begin{Bmatrix} .15 & .3339 \\ .1347 & .2491 \end{Bmatrix}_S$
0.875	.13799	.16000	.20420	.26278	.31696	.36715	.42009	.47373	1.04850	$\begin{Bmatrix} .125 & .3021 \\ .1057 & .2443 \end{Bmatrix}_S$
0.9	.12018	.15273	.20705	.26407	.31452	.36636	.42113	.47430	1.05148	$\begin{Bmatrix} .1 & .2604 & .4280 \\ .0777 & .2019 & .2203 \end{Bmatrix}_S$
0.925	.10747	.15289	.20977	.26175	.31391	.36799	.42056	.47316	1.05263	$\begin{Bmatrix} .075 & .2187 & .4019 \\ .0508 & .1733 & .2759 \end{Bmatrix}_S$
0.95	.10268	.15548	.20874	.26137	.31473	.36744	.42049	.47363	1.05281	$\begin{Bmatrix} .05 & .1622 & .3193 \\ .0260 & .1148 & .2232 \end{Bmatrix}_S$
1	.10261	.15559	.20858	.26156	.31455	.36753	.42053	.47350	1.05281	$\begin{Bmatrix} .0377 & .1243 & .2479 & .3864 \\ .0154 & .0728 & .1523 & .1866 \end{Bmatrix}_S$

Table A14. Case n = 16: minimax solution for quadratic loss

P	minimax rule d_m: $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	$d_m(5)$	$d_m(6)$	$d_m(7)$	minimax value (* 100)	least favorable symmetric prior
0.65	.35096	.35180	.35339	.35645	.36245	.37438	.39787	.43999	0.74028	$\begin{Bmatrix} .35 \\ .4126 \end{Bmatrix}_S$
0.7	.30108	.30251	.30578	.31302	.32817	.35628	.39897	.44953	0.85325	$\begin{Bmatrix} .3 \\ .3138 \end{Bmatrix}_S$
0.75	.25157	.25382	.25929	.27209	.29941	.34688	.40613	.45861	0.92517	$\begin{Bmatrix} .25 & .4452 \\ .2547 & .2453 \end{Bmatrix}_S$
0.8	.20345	.20893	.22222	.25057	.29777	.35333	.40373	.45070	0.96861	$\begin{Bmatrix} .2 & .3986 \\ .1854 & .3146 \end{Bmatrix}_S$
0.825	.17948	.18715	.20616	.24506	.30140	.35593	.40097	.44687	0.98174	$\begin{Bmatrix} .175 & .3719 \\ .1598 & .2876 \end{Bmatrix}_S$
0.85	.15784	.17045	.19864	.24674	.30295	.35320	.39964	.44844	0.99045	$\begin{Bmatrix} .15 & .3276 \\ .1268 & .2445 \end{Bmatrix}_S$
0.875	.13636	.15568	.19559	.25087	.30277	.34948	.39928	.45096	0.99612	$\begin{Bmatrix} .125 & .3002 & .4996 \\ .1002 & .2449 & .1550 \end{Bmatrix}_S$
0.9	.11880	.14812	.19786	.25220	.30058	.34894	.40002	.45080	0.99878	$\begin{Bmatrix} .1 & .2536 & .4201 \\ .0717 & .1901 & .2382 \end{Bmatrix}_S$
0.925	.10537	.14724	.20101	.25047	.29922	.35030	.40035	.44955	0.99982	$\begin{Bmatrix} .075 & .2125 & .3771 \\ .0464 & .1588 & .2247 \end{Bmatrix}_S$
0.95	.10023	.14973	.20029	.24978	.30010	.35002	.39992	.45007	1.00000	$\begin{Bmatrix} .05 & .1551 & .2933 & .4347 \\ .0230 & .0990 & .1807 & .1973 \end{Bmatrix}_S$
1	.10000	.15000	.20000	.24999	.30001	.34998	.40002	.44999	1.00000	$\begin{Bmatrix} .0391 & .1263 & .2496 & .3874 \\ .0146 & .0713 & .1524 & .1883 \end{Bmatrix}_S$

Appendix B

This appendix contains eight tables, presenting minimax estimators for the following estimation problem (Θ,L,X): parameter space Θ equals $[1-P,P]$ with $\frac{1}{2} < P \leq 1$, $X \in B(n,\theta)$ and the loss function is given by

$$L(\theta,a) = (\theta-a)^2/\{\theta(1-\theta)\}$$

(So the loss function constitutes the only difference with the situation of Appendix A.)

Each table successively presents for a certain n ($3 \leq n \leq 10$) and several values of P the following elements:

(i) the (symmetric) minimax estimator d_m, determined by the estimates $d_m(i)$ for $i = 0,1,\ldots,\mathrm{ent}(n/2)$;

(ii) the minimax value $\max_\tau r_\tau$;

(iii) the (discrete) symmetric least favorable prior distribution with the minimum number of steps - determined by (7.6).

Tabel B1. Case n = 3: minimax solution for weighted quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	minimax value (* 100)	symmetric least favorable prior
0.6	0.44571	0.48000	0.03703	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.39051	0.45500	0.07610	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.32919	0.42000	0.11987	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.26786	0.37500	0.16071	$\begin{Bmatrix}0.25\\0.5\end{Bmatrix}_S$
0.8	0.20923	0.32000	0.19052	$\begin{Bmatrix}0.2\\0.5\end{Bmatrix}_S$
0.85	0.17424	0.31204	0.21210	$\begin{Bmatrix}0.15\\0.3836\end{Bmatrix}_S$
0.9	0.13295	0.31219	0.24055	$\begin{Bmatrix}0.1\\0.2935\end{Bmatrix}_S$
0.95	0.07830	0.31641	0.27895	$\begin{Bmatrix}0.05\\0.2267\end{Bmatrix}_S$
1	0.00000	0.33333	0.33333	$\begin{Bmatrix}0\\0.1667\end{Bmatrix}_S$

<u>Table B2</u>. Case n = 4: minimax solution for weighted quadratic loss

P	minimax rule d_m		minimax	symmetric least
	$d_m(0)$	$d_m(1)$	value (* 100)	favorable prior
0.6	0.43299	0.46154	0.03569	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.37326	0.41743	0.07038	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.31305	0.36207	0.10503	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.25610	0.30000	0.13171	$\begin{Bmatrix}0.25\\0.5\end{Bmatrix}_S$
0.8	0.21346	0.26915	0.14763	$\begin{Bmatrix}0.2\\0.4168\end{Bmatrix}_S$
0.85	0.17364	0.26291	0.16568	$\begin{Bmatrix}0.15\\0.3171\end{Bmatrix}_S$
0.9	0.12562	0.25696	0.18822	$\begin{Bmatrix}0.1\\0.2508\end{Bmatrix}_S$
0.95	0.06836	0.25216	0.21600	$\begin{Bmatrix}0.05\\0.2034\end{Bmatrix}_S$
1	0.00000	0.25000	0.25000	$\begin{Bmatrix}0\\0.1667\end{Bmatrix}_S$

<u>Table B3</u>. Case n = 5: minimax solution for weighted quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	minimax value (* 100)	symmetric least favorable prior
0.6	0.42327	0.44571	0.48000	0.03442	$\begin{Bmatrix} 0.4 \\ 0.5 \end{Bmatrix}_S$
0.65	0.36299	0.39051	0.45500	0.06526	$\begin{Bmatrix} 0.35 \\ 0.5 \end{Bmatrix}_S$
0.7	0.30570	0.32919	0.42000	0.09241	$\begin{Bmatrix} 0.3 \\ 0.5 \end{Bmatrix}_S$
0.75	0.25585	0.27790	0.38693	0.10853	$\begin{Bmatrix} 0.25 \\ 0.4633 \end{Bmatrix}_S$
0.8	0.21540	0.25873	0.39200	0.12217	$\begin{Bmatrix} 0.2 \\ 0.3505 \end{Bmatrix}_S$
0.85	0.16851	0.23630	0.40258	0.13812	$\begin{Bmatrix} 0.15 \\ 0.2821 \end{Bmatrix}_S$
0.9	0.11652	0.21239	0.42268	0.15511	$\begin{Bmatrix} 0.1 \\ 0.2348 \end{Bmatrix}_S$
0.95	0.06661	0.19807	0.43459	0.17190	$\begin{Bmatrix} 0.05 & 0.3784 \\ 0.1715 & 0.3285 \end{Bmatrix}_S$
1	0.00000	0.20000	0.40000	0.20000	$\begin{Bmatrix} 0 & 0.2764 \\ 0.0833 & 0.4167 \end{Bmatrix}_S$

Table B4. Case n = 6: minimax solution for weighted quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	minimax value (* 100)	symmetric least favorable prior
0.6	0.41614	0.43299	0.46154	0.03323	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.35714	0.37326	0.41743	0.06064	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.30246	0.31305	0.36207	0.08168	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.25802	0.27624	0.33949	0.09318	$\begin{Bmatrix}0.25\\0.4063\end{Bmatrix}_S$
0.8	0.21285	0.24695	0.34114	0.10565	$\begin{Bmatrix}0.2\\0.3165\end{Bmatrix}_S$
0.85	0.16276	0.21220	0.34887	0.11916	$\begin{Bmatrix}0.15\\0.2640\end{Bmatrix}_S$
0.9	0.11185	0.17755	0.36463	0.13119	$\begin{Bmatrix}0.1 & 0.4351\\0.2206 & 0.2794\end{Bmatrix}_S$
0.95	0.06726	0.16838	0.35903	0.14429	$\begin{Bmatrix}0.05 & 0.3352\\0.1436 & 0.3564\end{Bmatrix}_S$
1	0.00000	0.16667	0.33333	0.16667	$\begin{Bmatrix}0 & 0.2764\\0.0833 & 0.4167\end{Bmatrix}_S$

Table B5. Case n = 7: minimax solution for weighted quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	minimax value (* 100)	symmetric least favorable prior
0.6	0.41106	0.42327	0.44571	0.48000	0.03210	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.35389	0.36299	0.39051	0.45500	0.05645	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.30106	0.30570	0.32919	0.42000	0.07237	$\begin{Bmatrix}0.3\\0.5\end{Bmatrix}_S$
0.75	0.25789	0.27347	0.31788	0.42382	0.08234	$\begin{Bmatrix}0.25\\0.3675\end{Bmatrix}_S$
0.8	0.20963	0.23550	0.30876	0.43312	0.09366	$\begin{Bmatrix}0.2\\0.2954\end{Bmatrix}_S$
0.85	0.15826	0.19224	0.30457	0.44827	0.10472	$\begin{Bmatrix}0.15\\0.2536\end{Bmatrix}_S$
0.9	0.11231	0.16315	0.30696	0.44925	0.11389	$\begin{Bmatrix}0.1 & 0.3851\\0.1925 & 0.3075\end{Bmatrix}_S$
0.95	0.06495	0.14806	0.30924	0.43235	0.12493	$\begin{Bmatrix}0.05 & 0.3238\\0.1302 & 0.3698\end{Bmatrix}_S$
1	0.00000	0.14286	0.28571	0.42857	0.14286	$\begin{Bmatrix}0 & 0.1727\\0.0500 & 0.2722\end{Bmatrix}_S$

Table B6. Case n = 8: minimax solution for weighted quadratic loss

P	minimax rule d_m				minimax	symmetric least
	$d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	value (* 100)	favorable prior
0.6	0.40751	0.41614	0.43299	0.46154	0.03102	$\begin{Bmatrix} 0.4 \\ 0.5 \end{Bmatrix}_S$
0.65	0.35211	0.35714	0.37326	0.41743	0.05263	$\begin{Bmatrix} 0.35 \\ 0.5 \end{Bmatrix}_S$
0.7	0.30255	0.30721	0.32295	0.37617	0.06477	$\begin{Bmatrix} 0.3 \\ 0.4574 \end{Bmatrix}_S$
0.75	0.25670	0.26949	0.30363	0.37964	0.07423	$\begin{Bmatrix} 0.25 \\ 0.3412 \end{Bmatrix}_S$
0.8	0.20679	0.22552	0.28247	0.38935	0.08436	$\begin{Bmatrix} 0.2 \\ 0.2816 \end{Bmatrix}_S$
0.85	0.15604	0.17943	0.26498	0.40644	0.09304	$\begin{Bmatrix} 0.15 & 0.4573 \\ 0.2422 & 0.2578 \end{Bmatrix}_S$
0.9	0.11155	0.15314	0.26732	0.39804	0.10098	$\begin{Bmatrix} 0.1 & 0.3662 \\ 0.1732 & 0.3268 \end{Bmatrix}_S$
0.95	0.06197	0.13052	0.27772	0.37700	0.11012	$\begin{Bmatrix} 0.05 & 0.3198 \\ 0.1231 & 0.3769 \end{Bmatrix}_S$
1	0.00000	0.12500	0.25000	0.37500	0.12500	$\begin{Bmatrix} 0 & 0.1727 \\ 0.0500 & 0.2722 \end{Bmatrix}_S$

<u>Table B7</u>. Case n = 9: minimax solution for weighted quadratic loss

P	minimax rule d_m $d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	minimax value (* 100)	least favorable symmetric prior
0.6	0.40507	0.41106	0.42327	0.44571	0.48000	0.03000	$\begin{Bmatrix}0.4\\0.5\end{Bmatrix}_S$
0.65	0.35114	0.35389	0.36299	0.39051	0.45500	0.04914	$\begin{Bmatrix}0.35\\0.5\end{Bmatrix}_S$
0.7	0.30316	0.30780	0.32024	0.35493	0.43906	0.05891	$\begin{Bmatrix}0.3\\0.4218\end{Bmatrix}_S$
0.75	0.25529	0.26536	0.29212	0.35118	0.44580	0.06784	$\begin{Bmatrix}0.25\\0.3224\end{Bmatrix}_S$
0.8	0.20461	0.21764	0.26030	0.35330	0.45706	0.07677	$\begin{Bmatrix}0.2\\0.2722\end{Bmatrix}_S$
0.85	0.15653	0.17600	0.23967	0.35903	0.46197	0.08379	$\begin{Bmatrix}0.15 & 0.4144\\0.2189 & 0.2811\end{Bmatrix}_S$
0.9	0.10999	0.14426	0.23956	0.35595	0.45166	0.09090	$\begin{Bmatrix}0.1 & 0.3575\\0.1605 & 0.3395\end{Bmatrix}_S$
0.95	0.06122	0.11763	0.24782	0.34376	0.44044	0.09836	$\begin{Bmatrix}0.05 & 0.2824\\0.1124 & 0.2915\end{Bmatrix}_S$
1	0.00000	0.11111	0.22222	0.33333	0.44445	0.11111	$\begin{Bmatrix}0 & 0.1175 & 0.3574\\0.0333 & 0.1893 & 0.2774\end{Bmatrix}_S$

Table B8. Case n = 10: minimax solution for weighted quadratic loss

P	minimax rule d_m					minimax	least favorable
	$d_m(0)$	$d_m(1)$	$d_m(2)$	$d_m(3)$	$d_m(4)$	value (* 100)	symmetric prior
0.6	0.40341	0.40751	0.41614	0.43299	0.46154	0.02903	$\begin{Bmatrix} 0.4 \\ 0.5 \end{Bmatrix}_S$
0.65	0.35061	0.35211	0.35714	0.37326	0.41743	0.04592	$\begin{Bmatrix} 0.35 \\ 0.5 \end{Bmatrix}_S$
0.7	0.30313	0.30740	0.31772	0.34291	0.40058	0.05424	$\begin{Bmatrix} 0.3 \\ 0.3949 \end{Bmatrix}_S$
0.75	0.25399	0.26164	0.28232	0.32958	0.40820	0.06262	$\begin{Bmatrix} 0.25 \\ 0.3084 \end{Bmatrix}_S$
0.8	0.20305	0.21184	0.24244	0.32008	0.42379	0.07037	$\begin{Bmatrix} 0.2 \\ 0.2657 \end{Bmatrix}_S$
0.85	0.15630	0.17275	0.22280	0.32114	0.42273	0.07640	$\begin{Bmatrix} 0.15 & 0.3971 \\ 0.2014 & 0.2986 \end{Bmatrix}_S$
0.9	0.10818	0.13608	0.21768	0.32467	0.40849	0.08271	$\begin{Bmatrix} 0.1 & 0.3531 \\ 0.1519 & 0.3481 \end{Bmatrix}_S$
0.95	0.06069	0.10860	0.22201	0.31540	0.39997	0.08901	$\begin{Bmatrix} 0.05 & 0.2605 \\ 0.1028 & 0.2641 \end{Bmatrix}_S$
1	0.00000	0.10000	0.20001	0.29999	0.40001	0.10000	$\begin{Bmatrix} 0 & 0.1175 & 0.3574 \\ 0.0334 & 0.1893 & 0.2774 \end{Bmatrix}_S$

Samenvatting

Bij schattingsproblemen staat de onbekende waarde van een zekere grootheid (de parameter) centraal. Op basis van waarnemingen die afkomstig zijn van een toevalsexperiment probeert de statisticus deze waarde zo goed mogelijk te bepalen. Een nadere precisering hiervan leidt tot allerlei criteria voor schatters, waaronder zuiverheid, toelaatbaarheid en invariantie. In een of andere zin optimaal zijn verder de (veel gebruikte) Bayes en minimax schatters.

Het waardebereik van de parameter heet de parameterruimte; intuïtief is duidelijk dat doorgaans betere schattingen mogelijk zijn naarmate de parameterruimte kleiner is. In dit proefschrift worden zgn. afgeknotte parameterruimten bestudeerd, waarbij het theoretisch mogelijke waardebereik van de parameter op voorhand al is ingeperkt. Dit kan gebeurd zijn op grond van logische redeneringen of theoretische inzichten; een voorbeeld hiervan doet zich voor bij de interview-methode die in de sociale wetenschappen als 'randomized response'-techniek bekend staat. Het kan echter ook zijn dat praktische ervaringen of subjectieve overwegingen tot een dergelijke inperking hebben geleid. Zo valt op grond van symmetrie-overwegingen goed te verdedigen dat voor elke Nederlandse munt de kans op het gooien van 'kruis' ligt tussen 0.4 en 0.6.

De centrale vraag is nu hoe zo'n afknotting van de parameterruimte gebruikt kan worden om betere schatters te verkrijgen. Het antwoord hangt uiteraard af van de gehanteerde criteria; daarnaast is van belang welke type kansverdeling de waarnemingen vertonen. Dit laatste aspect ligt ten grondslag aan de tweedeling van dit boek: Deel 1 behandelt de theorie voor zo algemeen mogelijke kansverdelingen, terwijl in Deel 2 uitsluitend de binomiale verdeling bestudeerd wordt. Binnen elk deel komen de eerdergenoemde criteria voor schatters achtereenvolgens aan de orde.

Afgeknotte parameterruimten komen in de praktijk regelmatig voor. Daarnaast zijn zij uit wiskundig en uit statistisch oogpunt van belang. Enkele van de meest interessante resultaten uit dit proefschrift zijn de volgende.

(i) In vele gevallen bestaan er bij een afgeknotte parameterruimte geen zuivere schatters. Dit verschijnsel maakt het criterium van zuiverheid onbruikbaar, zodat de nadruk noodgedwongen op de overige criteria valt.

(ii) Schatters die waarden aannemen dichtbij de rand van de afgeknotte parameterruimte zijn veelal ontoelaatbaar. Daardoor wordt niet alleen de beroemde schatter van maximale aannemelijkheid buiten spel gezet, maar ook bijna elke numerieke methode die momenteel bij afgeknotte parameterruimten wordt gebruikt.

(iii) Minimax schatters vertonen een zeer onregelmatig en soms verrassend gedrag. Omdat ze bovendien moeilijk te bepalen zijn, wordt in Deel 2 veel aandacht besteed aan de berekening van minimax schatters bij een binomiale verdeling.

In het algemeen worden in dit proefschrift bestaande schattingsmethoden toegespitst op het specifieke geval van een afgeknotte parameterruimte. Hoofdstuk 3 vormt hierop enigszins een uitzondering. De daar gepresenteerde theorie, die bovengenoemd punt (ii) combineert met invariantie van schatters, leidt ook voor klassieke, niet-afgeknotte parameterruimten tot nieuwe resultaten.

References

AKER, A.W. VAN DEN 1980, Een Algol 68 procedure voor niet-lineaire programmering (in Dutch), Report RC.8065, Tilburg University.

ALBERS, W. 1982, Simple randomized response procedures with bounded respondent risk for quantitative data, Kwantitatieve Methoden 8, 35-46.

ANDERSON, H. 1977, Efficiency versus protection in a general randomized response model, Scand. J. Statist. 4, 11-19.

BARNDORFF-NIELSEN, O. 1978, Information and exponential families, Wiley, New York.

BELLHOUSE, D.R. 1980, Linear models for randomized response designs, JASA 75, 1001-1004.

BICKEL, P.J. 1981, Minimax estimation of the mean of a normal distribution when the parameter space is restricted, Ann. Statist. 9, 1301-1309.

BIRNBAUM, Z.W. 1942, An inequality for Mill's ratio, Ann. Math. Statist. 13, 245-246.

BLUM, J.R. & ROSENBLATT, J. 1967, On partial a priori information in statistical inference, Ann. Math. Statist. 38, 1671-1678.

BOURKE, P.D. 1982, Randomized response multivariate designs for categorical data, Commun. Statist. A11, 2889-2901.

BOYD, A.V. 1959, Inequalities for Mills' ratio, Reports of Statistical Application Research, JUSE 6, 44-46.

CASELLA, G. & STRAWDERMAN, W.E. 1981, Estimating a bounded normal mean, Ann. Statist. 9, 870-878.

CHANG, Y.-T. 1981, Stein-type estimators for parameters restricted by linear inequalities, Keio Science and Technology Reports 34, 83-95.

CLICKNER, R.P. & IGLEWICZ, B. 1976, Warner's randomized response technique: the two sensitive questions case, Proc. Soc. Science Section A.S.A., 260-263.

DAVIS, W.W. 1978, Bayesian analysis of the linear model subject to linear inequality constraints, JASA 73, 573-579.

DEGROOT, M.H. 1970, Optimal statistical decisions, McGraw-Hill, New York.

DEVORE, J.L. 1977, A note on the randomized response technique, Commun. Statist. A6, 1525-1529.

DIACONIS, P. & YLVISAKER, D. 1979, Conjugate priors for exponential families, Ann. Statist. 7, 269-281.

DOWNS, T., GILLILAND, D.C. & KATZ, L. 1978, Probability in a conquested election, The American Statistician 32, 122-125.

DRANE, W. 1975, Randomized response to more than one question, Proc. Soc. Science Section A.S.A., 395-397.

DYKSTRA, R.D. 1983, An algorithm for restricted least squares regression, JASA 78, 837-842.

FELLER, W. 1950, An introduction to probability theory and its applications I, Wiley, New York.

FERGUSON, T.S. 1967, Mathematical statistics, a decision theoretic approach, Academic Press, New York.

FLIGNER, M.A., POLICELLO, G.E. & SINGH, J. 1977, A comparison of two randomized response survey methods with consideration for the level of respondent protection, Comm. Statist. A6, 1511-1524.

GOODSTADT, M.S. & GRUSON, V. 1975, The randomized response technique: a test on drug use, JASA 70, 814-818.

GORDON, R.D. 1941, Values of Mill's ratio of area to bounding ordinate and of the normal probability integral for large values of the argument, Ann. Math. Statist. 12, 364-366.

GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R. & HORVITZ, D.G. 1969, The unrelated question randomized response model: theoretical framework, JASA 64, 520-539.

HODGES, J.L. & LEHMANN, E.L. 1952, The use of previous experience in reaching statistical decisions, Ann. Math. Statist. 23, 396-407.

HORVITZ, D.G., SHAH, B.V. & SIMMONS, W.R. 1967, The unrelated question randomized response model, Soc. Stat. Section of the Am. Stat. Ass. 65-72.

HORVITZ, D.G., GREENBERG, B.G. & ABERNATHY, J.R. 1976, Randomized response: a data-gathering device for sensitive questions, Int. Stat. Rev. 44, 181-196.

JACKSON, D.A., O'DONOVAN, T.M., ZIMMER, W.J. & DEELY, J.J. 1970, G_2-minimax estimators in the exponential family, Biometrika 57, 439-443.

JOHNSON, N.L. & KOTZ, S. 1970, Continuous univariate distributions - 2, Houghton Mifflin Company, Boston.

KARLIN, S. & SHAPLEY, L.S. 1953, Geometry of moment spaces, American Math. Soc., Providence.

KARLIN, S. 1958, Admissibility for estimation with quadratic loss, Ann. Math. Statist. 29, 406-436.

KARLIN, S. & STUDDEN, W.J. 1966, Tchebycheff systems: with applications in analysis and statistics, Interscience, New York.

KATZ, M.W. 1961, Admissible and minimax estimates of parameters in truncated spaces, Ann. Math. Statist. 32, 136-142.

KINGMAN, J.F.C. & TAYLOR, S.J. 1966, Introduction to measure and probability, Cambridge University Press.

KROTKI, K.J. & FOX, B. 1974, The randomized response technique, the interview and self-administered questionnaire: an empirical comparison of fertility reports, Proc. Social Science Section A.S.A.

KROTKI, K.J. & McDANIEL, S.A. 1975, Three estimates of illegal abortions in Alberta, Canada: survey, mail-back questionnaire and randomized response technique, Proc. 40th session I.S.I. 46/4, 67-69.

LANKE, J. 1975, On the choice of the unrelated question in Simmons' version of randomized response, JASA 70, 80-83.

LANKE, J. 1976, On the degree of protection in randomized interviews, Int. Stat. Rev. 44, 197-203.

LAPLACE, P.S. 1812, Théorie analytique des probabilités, Gauthier Villars, Paris.

LEE, T.C., JUDGE, G.G. & ZELLNER, A. 1977, Estimating the parameters of the Markov probability model from aggregate time series data, North Holland Publ. Company, Amsterdam.

LEHMANN, E.L. 1959, Testing statistical hypotheses, Wiley, New York.

LEHMANN, E.L. 1983, Theory of point estimation, Wiley, New York.

LEYSIEFFER, F.W. & WARNER, S.L. 1976, Respondent jeopardy and optimal designs in randomized response models, JASA 71, 649-656.

LIEW, C.K. 1976, Inequality constraint least-squares estimation, JASA 71, 746-751.

LOCANDER, W., SUDMAN, S. & BRADBURN, N. 1976, An investigation of interview method, threat and response distortion, JASA 71, 269-275.
MILLS, J.P. 1926, Table of ratio: area to bounding ordinate, for any proportion of the normal curve, Biometrika 18, 395-400.
MISES, R. VON 1964, Mathematical theory of probability and statistics, Academic Press, New York.
MOORS, J.J.A. 1971, Optimization of the unrelated question randomized response model, JASA 66, 627-629.
MOORS, J.J.A. 1973, On the absolute moments of a normally distributed random variable, Report E.I.T. 39, Tilburg University.
MOORS, J.J.A. 1977, Estimation in truncated parameter spaces, Ter Discussie 77.052, Tilburg University.
MOORS, J.J.A. 1979, Momentproblems and truncated binomial estimation, Ter Discussie 79.092, Tilburg University.
MOORS, J.J.A. 1981a, Inadmissibility of linearly invariant estimators in truncated parameter spaces, Ter Discussie 81.01, Tilburg University.
MOORS, J.J.A. 1981b, Inadmissibility of linearly invariant estimators in truncated parameter spaces, JASA 76, 910-915.
NELSON, W. 1966, Minimax solution of statistical decision problems by iteration, Ann. Math. Statist. 37, 1643-1657.
NOORBALOOCHI, S. & MEEDEN, G. 1983, Unbiasedness as the dual of being Bayes, JASA 78, 619-623.
POLLOCK, K.H. & BEK, Y. 1976, A comparison of three randomized response models for quantitative data, JASA 71, 884-886.
POOLE, W.K. 1974, Estimation of the distribution function of a continuous type random variable through randomized response, JASA 69, 1002-1005.
POWELL, M.J.D. 1978, A fast algorithm for nonlinearly constraint optimization calculations, Numerical Analysis, Dundee 1977 (G.A. Watson, ed.), Lecture notes in Math. no. 630, Springer, Berlin.
RAFSKY, L.C. 1976, Inadmissibility of the observed proportion in large samples drawn with replacement, Report Bell Laboratories.
RAGHAVARAO, D. 1978, On an estimation problem in Warner's randomized response technique, Biometrics 34, 87-90.
RAIFFA, H. & SCHLAIFER, R. 1961, Applied statistical decision theory, M.I.T. Press, Cambridge.
SAMPFORD, M.R. 1953, Some inequalities on Mill's ratio and related functions, Ann. Math. Statist. 24, 130-132.
SCHAFER, R.E. 1976, Modified estimators for the binomial parameter, The American Statistician 30, 98-100.
SHIMIZU, I.M. & BONHAM, G.S. 1978, Randomized response technique in a national survey, JASA 73, 35-39.
SHOHAT, J.A. & TAMARKIN, T.D. 1943, The problem of moments, American Math. Soc., New York.
SKIBINSKY, M. & COTE, L. 1963, On the admissibility of some standard estimates in the presence of prior information, Ann. of Math. Statist. 34, 539-548.
STEINHAUS, H. 1957, The problem of estimation, Ann. of Math. Statist. 28, 633-648.
TAMHANE, A.C. 1981, Randomized response technique for multiple sensitive attributes, JASA 76, 916-923.
TOUTENBURG, H. 1982, Prior information in linear models, Wiley, New York.

TUKEY, J. 1977, Exploratory data analysis, Addison-Wesley, Reading, Mass.

VERDOOREN, L.R. 1976, Loten bij delicate vragen; een overzicht van 'randomized response'-technieken (in Dutch), Statistica Neerlandica 30, 7-24.

WALD, A. 1939, Limits of a distribution function determined by absolute moments and inequalities satisfied by absolute moments, Trans. of the Am. Math. Soc. 46, 280-306.

WARNER, S.L. 1965, Randomized response: a survey technique for eliminating evasive answer bias, JASA 60, 63-69.

WARNER, S.L. 1971, The linear randomized response model, JASA 66, 884-888.

WIDDER, D.V. 1971, An introduction to transform theory, Academic Press, New York.

WIDDER, D.V. 1972, The Laplace transform, Princeton Univ. Press, Princeton.

WINKLER, R.L. & FRANKLIN, L.A. 1979, Warner's randomized response model: a Bayesian approach, JASA 74, 207-214.

ZAANEN, A.C. 1967, Integration, North Holland, Amsterdam.

ZACKS, S. 1971, The theory of statistical inference, Wiley, New York.

ZEYTINOGLU, M. & MINTZ, M. 1984, Optimal fixed size confidence procedures for a restricted parameter space, Ann. Statist. 12, 945-957.

Index

STELLINGEN
bij het proefschrift
ESTIMATION IN TRUNCATED PARAMETER SPACES
van J.J.A. Moors

1. In vele gevallen corresponderen schattingen op de rand van een afgeknotte parameterruimte noodzakelijkerwijs met een ontoelaatbare schatter. Daarom zijn maximaal aannemelijke schatters bij afgeknotte parameterruimten vaak ontoelaatbaar.

2. Het is technisch onmogelijk een geldstuk te vervaardigen
   (i) waarvoor de kans op 'kruis' lager is dan 0.4 of hoger dan 0.6; en
   (ii) dat bij oppervlakkige inspectie niet van normale geldstukken te onderscheiden is.

   Ter staving: bij 6.000 worpen met een rijksdaalder waarvan de beeldzijde is uitgehold kwam 3.062 maal de beeldzijde boven.

3. Het toepassen van Bayesiaanse methoden door accountants kan zeer zinvol en verantwoord zijn.

4. Het is veelbetekenend dat de twee voornaamste werkzaamheden aan instellingen van wetenschappelijk onderwijs doorgaans worden aangeduid met 'onderzoekstaak' en 'onderwijslast'.

5. Noem een zin zelftellend als zij een complete opsomming geeft van de er in verwerkte letters zonder dat gebruik gemaakt wordt van cijfers. Definieer de lengte van een zin als het totale aantal er in verwerkte letters. Dan zijn de eenentwintig kortste zelftellende Nederlandse zinnen achtereenvolgens:

   (i) Vijf f's, vijf s's, vijf v's, vijf ij's. (25)
   (ii) Negen e's; zes n-en, s-en; twee g's, t's, w's, z-en. (29)
   (iii) Negen e's; zeven n-en, s-en; twee g's, t's, v's, w's, z's; 'n a. (34)
   (iv) - (xx) Als (iii), met de laatste letter gewijzigd.
   (xxi) Schat, acht a's, acht c's, acht h's, acht s's, acht t's, schat! (40)

   KOUSBROEK, R. 1984, De logologische ruimte, Meulenhoff, Amsterdam.

6. '... ver in hun purisme dat altsaxofonist Bart Boying en zanger Jaap de Kwaadsteniet enkele maanden geleden lange tijd op het eiland hebben doorgebracht om de orq...'.
   Bovenstaand fragment uit de Volkskrant van 15 april 1985 (p. 8, kolom 7) bevat het volledige alfabet en is negen letters korter dan het record van BATTUS 1981, p. 63.

   BATTUS 1981, Opperlandse taal- en letterkunde, Querido, Amsterdam.

7. Bezuinigingen bij het wetenschappelijk onderwijs dienen te beginnen bij de folklore rond promoties.

8. De officiële eliminatie van de slotletter 's' in Nederlandse woorden is zo tijdrovend dat ik op zijn vroegst in 2003 door het leven zal kunnen gaan als Dr. Han Moor.